Die Insignifikanz signifikanter Unterschiede: der Genauigkeitsanspruch von PISA ist illusorisch

Wuttke, Joachim

Veröffentlichungsversion / Published Version
Sammelwerksbeitrag / collection article

Die Insignifikanz signifikanter Unterschiede: Der Genauigkeitsanspruch von PISA ist illusorisch

Joachim Wuttke

Inhaltsverzeichnis

Zur Online-Ausgabe im SSOAR (2013)

Der ab S. 5 folgende Aufsatz ist text-, aber nicht seitengleich mit der Buchausgabe von 2007. Die empfohlene Zitierweise ist daher

> Joachim Wuttke: *Die Insignifikanz signifikanter Unterschiede: Der Genauigkeitsanspruch von PISA ist illusorisch.* In Thomas Jahnke, Wolfram Meyerhöfer (Hrsg): *PISA & Co — Kritik eines Programms. Zweite Auflage.* Hildesheim: Franzbecker (2007). ISBN 978-3-88120-464-4. Zitiert nach der Online-Fassung, `http://nbn-resolving.de/urn:nbn:de:0168-ssoar-359057` (2013).

Jedoch empfehle ich ausdrücklich die Anschaffung des gedruckten Buchs: es kostet nur 16,80 € und enthält neun weitere Aufsätze, die das Phänomen PISA auf verschiedenen Ebenen untersuchen, von den ideologischen Grundlagen standardisierter Tests über die Testkonstrukte und ihre Operationalisierung bis hin zu den praktischen Folgen. Ich rate ausdrücklich ab von der Verwendung der ersten Auflage von 2006; die darin enthaltene Urfassung meines Aufsatzes unter dem Titel „Fehler, Verzerrungen, Unsicherheiten in der PISA-Auswertung“ ist in einigen Punkten überholt, wie im nachstehenden Vorwort zur zweiten Auflage und in Anhang A ausführlich dargelegt.

Ich verantworte diesen Aufsatz unter meiner Privatadresse (Sulzbacher Straße 10, 80803 München); er ist in meiner Freizeit entstanden und steht in keinem Zusammenhang mit meiner derzeitigen beruflichen Tätigkeit in der physikalischen Grundlagenforschung. Für Rückfragen und Diskussionen stehe ich unter `mail@joachimwuttke.de` gerne zur Verfügung.

Zum Thema PISA bin ich durch das Zusammenspiel dreier biographischer Faktoren gekommen: Berufserfahrung in statistischer Physik, Datenauswertung und wissenschaftliche Programmierung, Unterrichtserfahrung als Studienreferendar für Mathematik und Physik und drei Jahre Auslandserfahrung. In Frankreich lebend, war ich immer wieder beeindruckt, wie menschliches Denken von der Sprache gelenkt wird. Als Lehrer war ich verblüfft, wie fünfzehnjährige Gymnasiasten vor einer gut geübten Aufgabe kapitulieren, weil sie sich von einer marginalen Umformulierung in einer nebensächlichen Angabe irritieren lassen. Auf diesem Hintergrund war meine Ausgangsfrage an PISA: Wie können sprachlastige Testaufgaben sprach- und kulturübergreifend funktionieren?

In den offiziellen PISA-Berichten fand ich keine überzeugenden Antworten. Bei den Aufgabenübersetzungen war auf kulturelle Anpassung komplett verzichtet worden; nicht einmal Eigennamen wurden ausgetauscht. Das Aufgabenformat, mit einem hohen Anteil an Multiple-Choice-Fragen, war deutschen Schülern bis dato unbekannt. Und in der Auswertung wurde nichts getan, um sprach- und kulturbedingte Verzerrungen mathematisch zu erfassen und auszugleichen.

Daraufhin lud ich mir den kompletten Datensatz von PISA 2000 und 2003 herunter und begann mit eigenen Analysen. Würden sich sprachliche und kulturelle Einflüsse statistisch nachweisen lassen? Die Antwort findet sich in den

Abschnitten 4.6 bis 4.9, die die Keimzelle des ganzen Aufsatzes bilden: Ja, die PISA-Rohdaten belegen, dass viele deutsche Schüler mit dem Multiple-Choice-Format nicht zurechtkommen. Dass Leseaufgaben nicht nur Lesefähigkeit, sondern kulturell geprägtes Weltwissen testen. Dass Aufgabenschwierigkeiten zwischen verschiedenen Ländern umso stärker variieren, je unähnlicher die Sprachen sind.

Um mit den Rohdaten arbeiten zu können, musste ich mich in die umfangreiche, an manchen Stellen zwar lückenhafte, insgesamt aber vorbildliche technische Dokumentation von PISA einlesen. Dabei kam ich aus dem Staunen nicht heraus: Bei der Stichprobenziehung und Testdurchführung gibt es eine ganze Reihe von Detailproblemen, mit denen die Teilnehmerstaaten uneinheitlich umgegangen sind. Kapitel 2 gibt einen systematischen Überblick über diese Inkonsistenzen. Ein Highlight: Die Menschenfreundlichkeit des hochgelobten finnischen Schulsystems geht so weit, dass Legastheniker von der Testung ausgeschlossen werden. In Deutschland hingegen hat man ein eigenes Testheft entwickelt, um auch Sonderschüler einzubeziehen.

Aus diesen Erkenntnissen ergab sich die nächste Frage: Lässt sich quantitativ abschätzen, wie stark sich derlei Verzerrungen auf das politisch so wirkungsmächtige Hauptresultat von PISA, die Nationenrangliste, auswirken? Dazu war es nötig, detailliert nachzuvollziehen, wie in PISA Rohpunkte in Kompetenzkennwerte umgerechnet werden. Wegen Lücken in der technischen Dokumentation erwies sich diese Rekonstruktion als unerwartet schwierig; sie nimmt das ganze Kapitel 3 ein. Für Leser, die sich die mathematischen Details ersparen möchten, hier eine extreme Kurzfassung: Im Prinzip wird Leistung in PISA ganz einfach als Anzahl richtig beantworteter Teilaufgaben gemessen. Da aber verschiedene Schüler verschiedene Aufgabenhefte bearbeiten, kann man die Lösungspunkte nicht direkt miteinander vergleichen und rechnet sie deshalb unter der Annahme eines psychologisch-demographischen Modells in eine „Kompetenz"-Skala um, was approximativ auf den in Abbildung 4 gezeigten Zusammenhang zwischen dem Prozentsatz gelöster Teilaufgaben und dem zugeschriebenen Kompetenzwert führt. Im Ergebnis finde ich, dass PISA-Ländermittelwerte mit systematischen Unsicherheiten behaftet sind, die deutlich größer sind als die offiziell angegebenen, rein stochastischen Standardabweichungen (Abschnitte 6.2 und 6.3).

Die der Kompetenzskalierung zugrundegelegten Modellannahmen kann man empirisch überprüfen. Es zeigt sich, dass sie bei etlichen PISA-Aufgaben massiv verletzt sind (Abschnitte 4.2 bis 4.5). Die eindimensionale Erklärung der empirischen Schülerleistungen durch eine fachbezogene „Kompetenz" wird den Daten nicht gerecht. Vielmehr zeigt sich, dass auch die Vertrautheit mit dem Testformat einen quantifizierbaren und erheblichen Einfluss auf die erreichte Punktzahl hat.

Nach dem Länderranking war das in Deutschland meistbeachtete PISA-Ergebnis der Zusammenhang zwischen Testleistung („Bildungserfolg“) und sozialem Hintergrund. Auch hier sind nach gründlicher Lektüre der PISA-Berichte und einem eigenen Blick in die Daten mehr Fragezeichen als Ausrufezeichen zu machen (Kapitel 5).

Eine zwanzigseitige Zusammenfassung dieses Aufsatzes ist auf Englisch in einem zweisprachigen Sammelband erschienen:

> J. Wuttke: *Uncertainties and Bias in PISA.* In Hopmann/Brinek/Retzl (eds.): *PISA zufolge PISA — PISA According to PISA. Hält PISA, was es verspricht? Does PISA Keep What It Promises?* Wien: Lit-Verlag (2007). ISBN 978-3-8258-0946-1. Online verfügbar unter `http://ssrn.com/abstract=1159042`.

Schließlich möchte ich auf einen Überblicksartikel hinweisen, in dem ich über die statistische Unsicherheit hinaus Diskussionsbedarf anmelde, von der Grundidee „evidenzbasierter Politik“ bis hin zur kurzschlüssigen öffentlichen Interpretation der PISA-Ergebnisse:

> J. Wuttke: *PISA: Nachträge zu einer nicht geführten Debatte.* Mitteilungen der Gesellschaft für Didaktik der Mathematik 87, 22-34 (2009). Online verfügbar unter `http://didaktik-der-mathematik.de/pdf/gdm-mitteilungen-87.pdf`.

Zur zweiten Auflage

Dies ist eine eingreifend überarbeitete und erheblich erweiterte Neufassung des Aufsatzes „Fehler, Verzerrungen, Unsicherheiten in der PISA-Auswertung“ (Wuttke 2006, im folgenden zitiert als W1) aus der ersten Auflage des vorliegenden Sammelbands. Um Verwechslungen vorzubeugen, ist auch der Titel neu.

In dieser Neufassung sind erste Reaktionen auf W1 berücksichtigt. Aus dem PISA-Konsortium und seinem Umfeld sind mir die folgenden Stellungnahmen bekannt geworden:

- Aus der internationalen Projektleitung: Schulz (2006). Schulz bezieht sich auf eine Vorstufe von W1, die ich im März 2006 an Mitglieder des PISA-Konsortiums geschickt hatte. Sein Schreiben hat mich leider erst nach Erscheinen von W1 erreicht.
- Aus der deutschen Projektleitung: Prenzel (2006) mit technischem Anhang Prenzel/Walter (2006). Prenzel verspricht eine Auseinandersetzung mit den Fragen „Wie solide ist PISA? oder Ist die Kritik von Joachim Wuttke begründet?“, geht aber nur auf drei von über zwanzig Kritikpunkten ein.
- Eine von der Kultusministerkonferenz angeforderte, eilig angefertigte „Stellungnahme“ von Köller (2006a). Köller leitet eine im Gefolge von PISA gegründete Testaufgabenredaktion („Institut für Qualitätsentwicklung im Bildungswesen“). In Anbetracht seiner wissenschaftlichen Sozialisierung hätte die KMK genauso gut einen der PISA-Autoren um ein Gutachten in eigener Sache bitten können. Lesenswert auch ein Interview, in dem sich Köller (2006b) Kritik mit Neid erklärt.
- Ein Interview mit dem PISA-Beiratsmitglied Klemm (2006).

Da diese Texte nur in flüchtiger Form im Internet veröffentlicht wurden, wäre es sinnvoll gewesen, sie durch Reproduktion in dieser zweiten Auflage bibliotheksfest zu machen. Leider haben Köller, Prenzel, Schulz und Walter den Abdruck nicht genehmigt; Köller nahm die Bitte um Abdruckgenehmigung vielmehr zum Anlass, sein Gutachten von seiner Institutswebsite zurückzuziehen.

Dass somit keine einzige Replik Anspruch auf wissenschaftliche Beachtlichkeit erhebt, hat auch ein Gutes: statt mich mit Errata, Addenda und Antikritiken auf W1 zu beziehen, kann ich mit der folgenden Synthese einen übersichtlicheren und vertieften Ausgangspunkt für die weitere Auseinandersetzung anbieten. Inwieweit ich dabei meine Kritik aus W1 aufrecht erhalte, ist in Anhang A dargelegt. Rückblicke auf die bisherige Debatte im Haupttext bezwecken vor allem, ein Standardargument für die Validität von PISA, die Expertise zahlreich zusammenwirkender Experten, näher zu beleuchten.

Wertvolle Anregungen entnehme ich auch zustimmenden Zuschriften, für die ich an dieser Stelle noch einmal herzlich danke. Korrespondenz und Diskussionen nach Vorträgen haben mir geholfen, präziser zu erkennen und zu benennen,

welche Schlussfolgerungen aus den numerischen Tatsachen zu ziehen sind. Ich danke insbesondere P. Bender und W. Meyerhöfer für mehrmaliges Gegenlesen sowie T. Hothorn, G. Kanig und R. V. Olsen für Literaturhinweise.

1 Einleitung

1.1 Untersuchungsziel

PISA ist ein genuin statistisches Unternehmen, das von Anfang an auf eine bestimmte, extrem reduktionistische Auswertung hin angelegt ist (vgl. Bottani/ Vrignaud S. 30). Unter Beteiligung Hunderttausender Schüler werden reichhaltige *Primärdaten* erhoben. Diese Primärdaten, einige zehn Megabyte, umfassen die Ergebnisse aus dem eigentlichen, *kognitiven* Leistungstest, sowie mit einem *Student Questionnaire* erhobene *Hintergrunddaten*. Die kognitiven Testergebnisse werden im wesentlichen nur dazu genutzt, Kennzahlen zu bestimmen, die als *Aufgabenschwierigkeiten* und *Schülerkompetenzen* gedeutet werden. Um den Text nicht mit Anführungszeichen zu überfrachten, soll dieser Sprachgebrauch hingenommen werden, obwohl sich im Ergebnis zeigen wird, dass die „Schwierigkeit“ von Aufgaben nicht ohne Willkür durch einen eindimensionalen Parameter ausgedrückt werden kann, und dass PISA anderes misst als nur „Kompetenz“ in bestimmten Fachgebieten.

Die Kompetenzwerte der einzelnen Probanden, immer noch Hunderttausende von Zahlen, könnte man als *Sekundärdaten* bezeichnen. Um zu interpretierbaren statistischen Aussagen zu kommen, werden diese Einzelergebnisse über Subpopulationen gemittelt, mit Hintergrunddaten verknüpft und zu *Tertiärdaten* aggregiert. Typische Tertiärdaten sind Mittelwerte, Standardabweichungen, Perzentilgrenzen, Korrelationskoeffizienten und Gradienten. Sie werden zumeist in Form von Nationen-Ranglisten veröffentlicht (OECD 2001, 2004a sowie zahlreiche Detailstudien und nationale Berichte). Auf diese Tertiärdaten gründen sich umfangreiche verbale Deutungen, die sich im besten Fall an den inhaltlichen Anforderungen der auf einer bestimmten „Kompetenzstufe“ lösbaren Aufgaben orientieren. Für Politik und Öffentlichkeit werden Kurzfassungen erstellt, wobei in jedem Staat andere Aspekte in den Vordergrund gerückt werden.[1]

Im folgenden soll untersucht werden, wie zuverlässig diese lange Schlusskette ist. Kritik wurde bisher vor allem an ihrem Ausgangspunkt geübt: an den Testaufgaben und deren theoretischer Fundierung. In diesem Buch äußern sich dazu aus verschiedenen Perspektiven Keitel, Jablonka, Bender und Gellert; an anderer Stelle unsere Koautoren Hagemeister (1999, zu TIMSS) und Meyerhöfer (2005). International und unter Einschluss von TIMSS könnte man bereits eine umfangreiche Bibliographie füllen. Komplementär dazu konzentriert sich dieser Aufsatz auf numerische Aspekte des Testablaufs und auf quantitativ greifbare

[1] Die Aufmerksamkeit konzentriert sich regelmäßig auf die fürs eigene Land ungünstigen Nachrichten. So löste PISA 2000 in Finnland einen Schock aus, weil große Leistungsunterschiede zwischen Jungen und Mädchen gefunden wurden. Erst als ein Pilgerstrom deutscher Bildungspolitiker einsetzte, gewöhnten sich die Finnen langsam daran, als Testsieger zu gelten (S. Hopmann, mündl. Mitteilung).

Unsicherheiten und Verzerrungen. Dabei gibt es gelegentlich Berührungspunkte mit der Aufgabenkritik: an Beispielen wie dem Missverstehen des Multiple-Choice-Formats im deutschen Sprachraum lässt sich zeigen, dass inhaltliche Mängel zu quantitativ bedeutsamen Verzerrungen führen.

Viele statistische Probleme von Schulleistungsuntersuchungen sind in Fachkreisen durchaus bekannt, zumeist jedoch nur isoliert erörtert worden. Eine Ausnahme ist die vernichtende Kritik, mit der Freudenthal (1975) die ersten großen internationalen Vergleichsstudien der Lächerlichkeit preisgab. Manche seiner Einwände sind auch heute noch aktuell, denn am Grundkonzept solcher Studien hat sich trotz zahlreicher technischer Verfeinerungen wenig geändert.[2]

Hier soll ein breiter Überblick über tatsächliche, wahrscheinliche und mögliche Verzerrungen in PISA gegeben werden. Teil 2 stellt die Repräsentativität der Stichprobe, Teil 4 die eindimensionale Bewertbarkeit der kognitiven Leistungen in Frage. Soweit möglich, werden Verzerrungen quantifiziert. Dazu ist es nötig, zu verstehen, wie Schülerleistungen in Punkte umgerechnet werden. Teil 3 soll diese „Skalierung“ erklären und ein Gefühl dafür geben, was es eigentlich bedeutet, wenn sich die Testleistung zweier Populationen um einen bestimmten Zahlenwert unterscheidet.

Ziel dieser Untersuchung ist es *nicht*, bestimmte für Deutschland gefundene Kernaussagen für falsch zu erklären. Ich vermute im Gegenteil, dass PISA in Deutschland nur deshalb so fulminant einschlagen konnte, weil die meistpublizierten Aussagen mit der Lebenserfahrung von Schülern, Eltern und Lehrern durchaus kompatibel sind. Gefragt werden soll nicht, ob diese Aussagen richtig oder falsch sind, sondern ob sie aus dem vorliegenden Datenmaterial abgeleitet werden können. Es soll gefragt werden, mit welchen Unsicherheiten dieses Material behaftet ist, und ob in Anbetracht dieser Unsicherheiten der für die Datenerhebung getriebene Aufwand nicht fehlgerichtet ist. Es soll gefragt werden, ob die zyklische Fortführung von PISA relevante neue Erkenntnisse bringen kann, und ob es zu rechtfertigen ist, dass zum Zweck der Messung von Trends eine Mehrheit der Testaufgaben geheim gehalten wird.

1.2 Maßstab: Der Genauigkeitsanspruch von PISA

Im folgenden soll die PISA-Studie an ihrem eigenen numerischen Genauigkeitsanspruch gemessen werden. Dieser Anspruch äußert sich am deutlichsten in den Ranglisten der nationalen Kompetenzmittelwerte (OECD 2004a, S. 59, 71, 81, 88, 92, 281, 294; OECD 2004b, S. 42). Zu den Mittelwerten werden Standardfehler angegeben, die in den meisten Fällen wenige Punkte auf der offiziellen

[2] Insofern ist es mehr als mutig, wenn sich die an PISA beteiligten Mathematikdidaktiker auf Freudenthals „realistic math education“ berufen. Wie vordergründig und inkonsistent der Realitätsbezug realer PISA-Aufgaben ist, hat Meyerhöfer (2005) detailliert dargelegt.

Skala (der mit Mittelwert 500 und Standardabweichung 100) betragen. Basierend auf diesen Standardfehlern wird in Tabellen angegeben, mit welchen Unsicherheiten die Rangplätze behaftet sind. Kreuztabellen geben an, zwischen welche Leistungsunterschiede zwischen Staaten statistisch signifikant sind.

Ein Extrembeispiel ist Island, dessen Leistungsmittelwert von 515 im Teilgebiet Mathematik mit einem Standardfehler von nur 1,4 behaftet ist. Mit 95 %iger Sicherheit nimmt Island unter 29 OECD-Staaten einen Rang zwischen dem 10ten und dem 13ten ein. Die isländischen Schülerleistungen sind signifikant schlechter als die australischen ($524 \pm 2{,}1$), aber signifikant besser als die schwedischen ($509 \pm 2{,}6$). Allerdings nimmt die Bonferroni-Korrektur, die 95 %ige Sicherheit nicht nur für einzelne Vergleiche, sondern für das Ranking als ganzes gewährleisten soll, dem Unterschied zwischen Island und Schweden die Signifikanz; signifikant ist dann erst der Unterschied zwischen Island und Deutschland ($503 \pm 3{,}3$).[3]

Ein entgegengesetztes Extrembeispiel ist das Leseergebnis für Österreich von $491 \pm 3{,}8$ Punkten, das jede Einstufung zwischen dem 12. und dem 21. OECD-Rangplatz zulässt; die Abstände zu Norwegen (500) und Italien (476) sind noch nicht Bonferroni-signifikant, sondern erst die zu Belgien (507) und Griechenland (472).

Je nach Vergleichspaar sind also Differenzen zwischen 9 (Australien – Island) und 19 (Österreich – Griechenland) Punkten nötig, um einen signifikanten Unterschied zwischen zwei Staaten zu begründen. Die solchen Aussagen zugrundeliegenden Standardfehler berücksichtigen jedoch nur bestimmte stochastische Unsicherheiten. Es liegt nahe, dass es in einer so komplexen Erhebung zahlreiche weitere Quellen von Ungenauigkeiten geben kann. Von der Auswahl und Übersetzung der Aufgaben über die Ziehung der Stichproben und die Durchführung in den Schulen bis hin zur Kodierung der Antworten und Aufbereitung der Daten hat jede Projektphase ihre eigenen Schwierigkeiten. Im folgenden soll geprüft werden, ob die Genauigkeit von PISA nicht eher durch systematische Fehler und Unsicherheiten als durch stochastische Standardfehler begrenzt wird.

Diesen Ansatz beurteilt Klemm (2006) in seiner kurzen Stellungnahme zu W1 so:

> Wuttke macht den fundamentalen Fehler, das Spiel der Ranglisten mitzuspielen. [...] Es gibt trotzdem eine ganze Reihe von Ländern, die unter ähnlichen Bedingungen leben und arbeiten wie wir und die bessere Leistungen bringen. Daran ist nicht zu rütteln. Die zentralen Aussagen von PISA bleiben: Deutschland

[3] Dieser Genauigkeitsanspruch ist so überzogen, dass er selbst Verteidigern des Unternehmens suspekt ist. Unter den fünfundzwanzig (!) Gründen, mit denen sich die *Zeitschrift für Pädagogik* einem Abdruck meiner ersten Ergebnisse entzog, war der erste und längste, dass obige 95 %-Aussagen nicht korrekt zitiert und „u. E. schlicht weg falsch“ seien. Der Gutachter hatte sich nicht die Mühe gemacht, in der angegebenen Quelle nachzuschlagen.

> dümpelt im Mittelfeld, nirgends sonst entscheidet die soziale Herkunft derart stark über den Bildungsstand, und Deutschland fördert Kinder mit Migrationshintergrund besonders schlecht.

Die *Beschreibung* meines Vorgehens als ein „Mitspielen“ ist treffend: um PISA immanent zu kritisieren, lasse ich mich auf gewisse Prämissen ein, aus denen sich unvermeidlich die Konstruierbarkeit von Ranglisten ergibt. Die *Bewertung* von Ranglisten als ein Spiel ist hingegen keine legitime Position für jemanden, der als Mitglied des wissenschaftlichen Beirats für PISA mitverantwortlich ist – umso mehr, als die Ranglisten keine Marginalie sind, sondern eine so zentrale Stellung haben, dass es das ganze Unternehmen ohne sie nicht gäbe. Einem Kritiker zum Vorwurf zu machen, er mache einen „fundamentalen Fehler“, wenn er ernstnimmt, was der Öffentlichkeit als ernsthafte Forschung verkauft wird, ist zynisch.

Klemms Position ist überdies inkonsistent: seine Behauptungen, Deutschland dümpele „im Mittelfeld“ und „eine ganze Reihe von Ländern“ mit ähnlichen Lebensbedingungen brächte „bessere Leistungen“ sind nichts anderes als Paraphrasen von Rangplätzen.[4] Und die Aussage, „nirgends sonst“ hinge die Testleistung („Bildungsstand“) so stark von der sozialen Herkunft ab, ist nichts anderes als die Angabe eines ganz bestimmten Rangs, nämlich des ersten in einer Rangliste einer nicht genau spezifizierten mathematischen Beziehung zwischen Kompetenzwerten und Hintergrunddaten.[5]

1.3 Datenbasis und Methodisches

Der Übersichtlichkeit halber beschränken sich die folgenden Analysen auf PISA 2003. Nationale Ergänzungsstudien bleiben unberücksichtigt.[6] Soweit die vier

[4] An denen im übrigen sehr wohl zu rütteln ist: wenn sich Deutschland im Lesetest aus PISA 2003 tatsächlich, wie der Interviewer suggerierte, um sechs Plätze verbesserte, dann brächten nur die Niederlande (vor Bonferroni-Korrektur), Schweden, Irland, Neuseeland, Australien, Kanada, Südkorea und Finnland signifikant bessere Leistungen. Von diesen Staaten haben lediglich die Niederlande und Schweden einen erheblichen Bevölkerungsanteil überwiegend gering qualifizierter Immigranten. Die „ganze Reihe von Ländern“ mit vergleichbaren Bedingungen und unzweifelhaft besseren Leistungen gibt es nicht.

[5] Inhaltlich ist Klemm auf dem Stand von PISA 2000: In PISA 2003 konnte der genannte erste Platz nicht reproduziert werden. Das Konsortium musste sogar vom sozialen Gradienten auf eine andere Kennzahl, die „Varianzaufklärung“, ausweichen, um für Deutschland wenigstens einen Wert zu finden, der signifikant schlechter als der OECD-Durchschnitt ist (Prenzel *et al.* 2004b, S. 248, 251). Das Motiv für diesen wissentlich herbeigeführten *publication bias* liegt auf der Hand: nur eine schlechte Nachricht ist eine gute Nachricht im Sinne der politischen Wirkungsabsicht.

[6] Eine eigenständige Analyse des deutschen Datensatzes ist kaum möglich, da die *public use files* aus Furcht vor unautorisierten Bundesländervergleichen keine Länder- und Schulkennungen enthalten; selbst die nur nach Vertraulichkeitsvereinbarung zugänglichen *scientific use files* sind unvollständig (Baumert/Artelt 2005).

Testgebiete (Lesen, Mathematik, Naturwissenschaften, Problemlösen) unterschieden werden müssen, wird als bevorzugtes Beispiel das schwerpunktmäßig untersuchte Gebiet Mathematik gewählt. Testergebnisse werden primär dem internationalen Ergebnisbericht (Learning for Tomorrow's World, OECD 2004a, im folgenden als LTW zitiert), dem Technischen Bericht (Technical Report, OECD 2005a, im folgenden als TR zitiert), sowie eigenen Auswertungen des internationalen Datensatzes (ACER 2005) entnommen.

Für die Rekonstruktion der Auswerteprozeduren erweist sich der Technische Bericht als unzureichend. Unentbehrlich, wenngleich ebenfalls schlecht geschrieben und fehlerhaft, sind einige Seiten im Handbuch zum Skalierungsprogramm ConQuest (Wu *et al.* 1998). Im Kern basiert die Skalierung der Aufgabenschwierigkeiten und Schülerkompetenzen auf der probabilistischen Testtheorie (*Item Response Theory*, IRT). Zu dieser gibt es zwar etliche Monographien; überwiegend sind das aber einführende Lehrwerke, die wenig Mathematik voraussetzen, Elementares in ermüdender Breite auswalzen und kaum zum Kernproblem der Parameterschätzung vordringen – mit Ausnahme von Baker (1992), der sich aber ebenfalls in kleinschrittiger Algebra verliert. Gehaltvoll und hilfreich ist ein Sammelband (Fischer/Molenaar 1995), den mir dankenswerterweise Prenzel (2006) in seiner Reaktion auf W1 empfohlen hat; in den OECD-Berichten wird er nicht zitiert.

Der internationale Datensatz besteht aus ASCII-Dateien, die sowohl Primärdaten als auch Skalierungsergebnisse enthalten. Wie die 2086 Byte pro Proband zu lesen sind, ist im Auswertehandbuch (*Data Analysis Manual*, OECD 2005b, im weiteren zitiert als DAM) beschrieben. Manche Angaben sind unklar oder unzureichend, aber über einen „helpdesk“ konnte ich Projektmitarbeiter erreichen, die e-Mail-Anfragen in der Regel zügig beantworteten – jedenfalls solange meine Fragen nicht auf ein gründliches Verständnis der Skalierung zielten.[7] Eigene Auswertungen erfordern vor allem einiges Umsortieren des Datensatzes und die manuelle Eingabe kleinerer Tabellen (Lösungscodes, Zusammensetzung der Testhefte u. a.). Daran lassen sich beliebige statistische Analysen anschließen.

Bei Mittelungen über den Datensatz lehne ich mich eng an die offizielle Auswertung an: in der Regel beziehe ich alle dreißig OECD-Staaten mit gleichem Gewicht ein; die Partnerstaaten lasse ich unberücksichtigt. Großbritannien[8] wurde wegen verfehlter Teilnahmequoren in sehr inkonsequenter Weise von

[7] Meine Fragen haben ACER zu einer Korrektur eines Codebooks (A. Berezmer, Mail vom 24. 1. 2005) und zur Herausgabe eines Erratums (P. McKelvie, Mail vom 9. 2. 2006) veranlasst, was darauf hindeutet, dass eine unabhängige Analyse der Originaldaten bisher nicht von vielen unternommen wurde.

[8] Gemeint ist hier und im folgenden stets das Vereinigte Königreich einschließlich Nordirland.

der offiziellen Auswertung ausgeschlossen: bei der Skalierung der Aufgabenschwierigkeiten und Schülerkompetenzen und bei der Berechnung von OECD-Mittelwerten wurden die britischen Daten noch einbezogen; nur in den Staaten-Ranglisten des Ergebnisberichts werden sie nicht aufgeführt (LTW, S. 33). Um meine Daten möglichst vergleichbar mit den offiziellen Berichten zu halten, beziehe ich Großbritannien durchgehend ein.

Auf inferenzstatistische Hypothesentests verzichte ich bewusst; ich bestreite die Logik des Konsortiums, in der noch so plausible Quellen von Verzerrungen erst dann zugegeben werden, wenn ihre statistische Signifikanz bewiesen werden kann (dazu Beispiele in 4.8 und 5.1).

1.4 Einmischung ohne Einblicke

Über W1 stand ohne weiteren Kommentar:

> One example for the need of mathematical literacy is the frequent demand for individuals to make judgements and to assess the accuracy of conclusions and claims in surveys and studies. Being able to judge the soundness of the claims from such arguments is, and increasingly will be, a critical aspect of being a responsible citizen.
> *The PISA 2003 Assessment Framework* (OECD 2003a, S. 27).

Dieses Zitat zeigt, wie unrealistisch das für PISA zentrale Konstrukt „literacy“ ist (vgl. Shamos 1995). Die Qualität einer statistischen Untersuchung zu beurteilen, kann man nicht ernsthaft von Fünfzehnjährigen erwarten. Nachdem mir mehrere Kultusministerien sowie der Generalsekretär der Kultusministerkonferenz erklärt haben, meine Statistik-Kritik inhaltlich nicht beurteilen zu können, scheint es, dass selbst für PISA zuständige Ministerialbeamte nicht über mathematische Grundbildung im Sinne der OECD verfügen.

Neben aller Ironie war das Motto auch programmatisch gemeint: mein Standpunkt ist nicht der eines Fachkollegen, der, wie es Prenzel (2006) bevorzugt hätte, „Optimierungen“ vorschlägt, sondern der eines Staatsbürgers, der sich ein Urteil über die Genauigkeit einer folgenreichen Studie gebildet hat und nun auf deren begrenzte Aussagekraft hinweist. Deshalb hätte sich Köller (2006a) die Mühe sparen können, mich – ohne auf den Kern meiner Argumentation einzugehen – als „Laien“ zu „entlarven“, der keine „Einblicke in die Szene“ hat. Eine solche Logik, die von sachlicher Kommunikation abhebt und allein auf Reputationswerte abstellt („Full Professor an der University of California, Berkeley, ein äußerst renommierter Professor“), setzt ein „gegenüber den strengen Wahrheitskriterien verringertes Anspruchsniveau“ (Luhmann 1974, S. 237) voraus; zugleich immunisiert sie perfekt gegen Kritik, denn die wird, in Anbetracht der inzestuösen Enge der „Szene“ (Beispiel: Köllers eigener Werdegang), immer nur von außen kommen.

Tatsächlich braucht es keine Spezialkenntnisse, sondern vor allem Geduld beim Lesen umfangreicher, nachlässig redigierter Berichte, um einen ersten Überblick über die Ungereimtheiten von PISA zu bekommen. Um Köllers Behauptung zu entkräften, das Skalierungsverfahren von PISA sei allgemein als nicht zu übertreffender „State of the Art“ anerkannt und durch „Hunderte von Aufsätzen“ gedeckt, benötigt man lediglich einen Bibliotheksausweis.

2 Wie repräsentativ ist PISA?

Schon bei der Definition der Grundgesamtheit (2.1) zeigt sich, dass der theoretische Anspruch von PISA, den „outcome“ von Schulsystemen zu messen, nicht erfüllbar ist. Aber auch wenn man die vorgegebene Zielpopulation hinnimmt, findet man, dass Uneinheitlichkeiten bei Stichprobenziehung (2.2 ff.), Teilnahmequoten (2.6 ff.) und Durchführung (2.9) mit dem Genauigkeitsanspruch der Studie nicht vereinbar sind. Einige Verzerrungen lassen sich größenordungsmäßig abschätzen; es zeigt sich, dass sie, gemessen am stochastischen Standardfehler, quantitativ bedeutsam sind.

2.1 Schuleinschreibungsquoten

Erklärtes Ziel von PISA ist es, durch internationalen Vergleich die leistungsmäßigen Ergebnisse der nationalen Bildungssysteme (*outcomes in terms of student achievements*, OECD 2003a, S. 6) zu beobachten. Als Grundgesamtheit wurden jedoch nicht die Absolventen gewählt, sondern fünfzehnjährige Schüler.[9] In diesem Alter haben viele Schüler noch mehrere Schuljahre vor sich; die Fähigkeit zum abstrakten Denken befindet sich mitten in der Entwicklung (z. B. Gräber/ Stork 1984, Carroll 1987). Deshalb ist von vornherein klar, dass PISA *nicht* Endergebnisse ganzer Bildungssysteme misst. Durch die bloße Behauptung, die von PISA erfassten Fertigkeiten spiegelten die Befähigung zu weiterem Lernen wider (OECD 2003a, S. 8), wird das nicht behoben.

Bestenfalls liefert PISA eine Momentaufnahme des Leistungsstandes im gewählten Alter. Doch in manchen Staaten geht ein nennenswerter Teil der Jugendlichen schon vor oder während dem fünfzehnten Lebensjahr von der Schule ab. In der Türkei beträgt die Einschreibungsquote im PISA-Jahrgang nur 54 %, in Mexiko 58 %. Auch in vielen Nicht-OECD-Partnerstaaten ist diese Quote so niedrig, dass die PISA-Ergebnisse nicht einmal näherungsweise für einen Altersjahrgang repräsentativ sind. Was haben solche Staaten davon, sich für PISA als „Ranking-Staffage“ (Meyerhöfer 2006a) zur Verfügung zu stellen? Schulz (2006, Punkt 1) bestätigt ausdrücklich, dass die Testung ganz auf entwickelte Länder ausgelegt ist.

In Teststatistiken erscheint ein Bildungssystem umso leistungsfähiger, je mehr Schwänzer und Abbrecher es produziert.[10] In Portugal gehen allein in den höchstens zwei Monaten, die zwischen der Stichprobenziehung und der Testung

[9] Die Schüler müssen mindestens im 7. Schuljahr sein. Diese Einschränkung dürfte in den meisten Staaten keine nennenswerte Auswirkung haben.

[10] So schon Freudenthal (1975, S. 151). In den USA verzerrt dieser Effekt nicht mehr bloß Daten, sondern Lebenschancen: es gibt Anzeichen, dass der Testdruck, der auf die einzelnen Schulen ausgeübt wird, diese dazu veranlasst, schwache Schüler herauszudrängen (Kohn 2000, S. 40 f.; Shriberg/Shriberg 2006).

liegen (P. McKelvie, Mail vom 23. 1. 2006), über fünf Prozent des Jahrgangs von der Schule ab, wodurch die Einschreibungsquote auf unter 86 % sinkt (TR, S. 168 f.). Wenn man auch Schulabgänger zum Ergebnis eines Bildungssystems zählt und konservativ abschätzt, dass diese im Mittel um nur eine Standardabweichung schwächer sind als die verbleibenden Schüler, dann überschätzt PISA die Leistung des portugiesischen Systems um mehr als 14 Punkte.

Darauf entgegnet Schulz (2006, Punkt 1):

> In most OECD countries, the vast majority of 15-year-olds are still enrolled in school [...] '15-year-olds enrolled in schools' is probably the closest one can get to the end of compulsory schooling across OECD countries.

Diese Argumentation ist typisch für Reaktionen des Konsortiums: Schulz lässt sich gar nicht darauf ein, dass ich die Validität von PISA *quantitativ*, in Relation zum numerischen Genauigkeitsanspruch der Studie, in Frage stelle. Formulierungen wie „vast majority" sind ungeeignet, diesen Anspruch zu verteidigen. Dass man wahrscheinlich nicht genauer messen kann, ist kein Argument für die Tauglichkeit einer Messung.

2.2 Stichprobenziehung: Inkonsistente Ausgangsdaten

Für PISA 2003 wurde eine Mindeststichprobengröße von 4500 Schülern gefordert und überall erreicht.[11] Mit Ausnahme weniger Staaten wurde die Schülerstichprobe in einem zweistufigen Verfahren gezogen: zunächst wurden Schulen ausgewählt; in Schulen, deren Jahrgangsstärke über einer Sollstichprobengröße n (in den meisten Staaten 35) lag, wurden anschließend n Schüler ausgewählt. Dieses Verfahren bringt mit sich, dass Probanden aus verschiedenen Schulen verschiedene statistische Gewichte zugeordnet werden müssen.[12] Die Gewichte können so weit auseinander liegen, dass sie durch einen willkürlichen *trimming factor* begrenzt werden müssen, damit nicht das Gesamtergebnis übermäßig von einzelnen Schülern abhängt (TR, S. 108 ff.).

Für die Ziehung der Schulstichprobe und für die Festlegung der Gewichte ist im Prinzip eine Urliste erforderlich, die für alle Schulen des Teilnehmerstaats aufführt, wieviele Schüler dem Testjahrgang angehören. Eine solche Liste ist in den wenigsten Staaten verfügbar. Sie durfte deshalb durch eines von vier Schätzverfahren ersetzt werden. In Griechenland fehlten die Voraussetzungen selbst für das gröbste Schätzverfahren, so dass alle Schulen gleichgewichtet werden mussten (TR, S. 52). Für Schweden wurde eine Einschreibungsquote von

[11] Außer in Island und Luxemburg, wo das Mögliche getan und ein ganzer Jahrgang getestet wurde (TR, S. 48, 173).

[12] Im folgenden nenne ich, ohne Garantie für Konsistenz, die ungewichteten Testteilnehmer „Probanden", die gewichteten Testteilnehmer „Schüler", weil sie repräsentativ für alle Schüler eines Landes sein sollen.

102,5 % registriert, für die Toskana von 107,7 %. Das lässt zumindest ahnen, wie inkonsistent mancherorts das verwendete Datenmaterial war (TR, S. 168, 183).

Auch die Einschreibungsquote von 100,000 % in den USA (TR, S. 168) ist vollkommen unglaubwürdig. Tatsächlich kennt das United States Department of Education diese Quote so schlecht, dass es zum Mittel einer Umfrage greifen musste, um herauszufinden, dass 2003 circa 2,2 % der „student population" zu Hause unterrichtet wurden (NCES 2006), also nicht in die PISA-Kategorie „eingeschriebene Schüler" (TR, S. 46) fielen.

2.3 Stichprobenziehung: Inkonsistente Stratifizierung

Im Alter von fünfzehn Jahren gehen nicht nur die ersten Jugendlichen von der Schule ab; in vielen Ländern findet ungefähr in diesem Alter ein Schulwechsel statt, und das Schulsystem fächert sich weit auf. Das Leistungsmittel in unterschiedlichen Schulformen liegt oft um mehr als eine Standardabweichung auseinander (2.7). Damit die PISA-Stichprobe repräsentativ für den Altersjahrgang ist, müssen die verschiedenen Schulformen im richtigen Verhältnis herangezogen und gewichtet werden. Schüler, die eine Klasse wiederholt haben, sind anders auf Schulformen verteilt als Gleichaltrige, die eine glatte Schullaufbahn hinter sich haben. In manchen Schulformen sind alle Fünfzehnjährigen Repetenten (Putz 2006), was zu sehr kleinen Stichproben führen kann, die korrekt zu berücksichtigen besonders schwierig ist („an administrativ burden", TR, S. 53). All dies sind potentielle Quellen systematischer Verzerrungen, die bei entsprechendem Interesse auch für gezielte Manipulationen genutzt werden können.[13]

Hauptziel der zyklischen Wiederholung von PISA ist es, zeitliche Veränderungen von Bildungssystemen zu beobachten. Im Vergleich von 2003 und 2000 war die Effektstärke mehrheitlich geringer als das rein stochastische Rauschen, in Österreich aber wurde eine auf 99 %-Niveau signifikante (LTW, S. 282) Abnahme der Leseleistung festgestellt. In der Öffentlichkeit wurde die Verschlechterung um neun OECD-Rangplätze als „Absturz" wahrgenommen und allen Ernstes mit dem zwischenzeitlich erfolgten Regierungswechsel in Verbindung gebracht.[14]

---

[13] Zu Manipulationsmöglichkeiten in *high-stakes tests* siehe Haladyna *et al.* (1991), Kraus (2005, S. 51) und Nichols/Berliner (2007, insbesondere Kap. 4: „States Cheat Too!"). PISA ist *low-stakes* für die Probanden, aber nicht immer für die Schulen (spätestens dann nicht, wenn Ergebnisse an die Öffentlichkeit gegeben werden: siehe zum Beispiel den Rummel um die Wiesbadener Helene-Lange-Schule), und bestimmt nicht für die Erziehungsministerien (wie zufällig ist es, dass sich unter den wenigen Schulen, die Hessen zur internationalen Stichprobe beiträgt, eine „Versuchsschule" und „UNESCO-Projektschule" befindet?).

[14] So noch am 24. 2. 2007 auf der Wikipedia-Seite über Elisabeth Gehrer. Dort wird auch behauptet, Gewerkschaft und Wirtschaftskammer hätten sich mit dem Ministerium zusammengetan, um künftig „mittels eines Prospekts die zu prüfenden Schüler auf die Tücken solcher Tests vorzubereiten." Wie 4.6 zeigen wird, könnte ein intelligent gemachter Prospekt gerade in Österreich viel bewirken.

Daraufhin veranlasste das Bildungsministerium eine Überprüfung. Neuwirth, Ponocny und Grossmann (2006a) stellten „bei näherer Betrachtung [...] schon bald Inkonsistenzen bei den österreichischen Daten“ fest (S. 11). Ihr Bericht mit dem unverfänglichen Titel „PISA 2000 und PISA 2003: Vertiefende Analysen und Beiträge zur Methodik“ ist nichts anderes als ein umfangreiches Erratum zu allen vorigen Berichten. Der vermeintliche Absturz erklärt sich damit, dass

> bei PISA 2000 die Berufsschulen deutlich unterrepräsentiert waren und daher die publizierten österreichischen Gesamtergebnisse 2000 besser waren, als es den tatsächlichen Verhältnissen entspricht [S. 13].

Zur falschen Gewichtung war es durch die Mischung inkonsistenter Schülerzahlen gekommen: in der einen Zahl waren sämtliche fünfzehnjährigen Berufsschüler berücksichtigt worden, in der anderen fehlten die, die im Testzeitraum eine Praxisphase hatten.

> Da die betroffene Gruppe vor allem aus männlichen Schülern besteht, kam es in weiterer Folge auch noch zu Verzerrungen bei allen Fragen, die das Geschlecht betreffen [S. 36].

Außerdem weisen Neuwirth *et al.* (S. 13 f., 52 ff.) auf Verzerrungen durch die Normierung auf ein bestimmtes Testheft hin (dazu unten 3.10).

Andreas Schleicher steuert zu dem Bericht ein Vorwort bei. Nach einer Seite mit Standardtextbausteinen geht er nur kurz auf den konkreten Inhalt des Berichts ein, ohne irgendeinen Fehler einzugestehen:

> Die österreichischen Stichproben erfüllen die strengen Kriterien der OECD erst seit PISA 2003, so dass die OECD bislang darauf hinweisen musste, dass eine Bewertung von Leistungsveränderungen seit PISA 2000 für Österreich nur eingeschränkt zulässig ist, ohne die potentiellen Verzerrungen jedoch quantifizieren zu können. Der vorliegende Band schließt diese wissenschaftliche Lücke und ermöglicht erstmals bereinigte Trendanalysen für Österreich [S. 10].

Allerdings muss die OECD ihren Hinweis, dass alle Trendaussagen für Österreich unter Vorbehalt standen, *sehr* kleingedruckt haben: mit Volltextsuche nach „Austria“ ist er im Ergebnisbericht (LTW) nicht zu finden. In die Öffentlichkeit ist er erst recht nicht transportiert worden. Warum haben die PISA-Verantwortlichen in ihren Aufsätzen, Interviews und Vorträgen nicht eindringlich auf die Unsicherheit der Stichprobe hingewiesen? Und was ist mit den Ergebnissen aus PISA 2000? Sind die seinerzeit unter Vorbehalt veröffentlicht worden? Warum hat das niemand mitbekommen? Die Öffentlichkeit ist mindestens durch unterlassene Aufklärung getäuscht worden.

Unregelmäßigkeiten bei der Stichprobenziehung dürften auch Anteil an den hervorragenden Ergebnissen Südtirols haben (Putz 2005, 2006). Obwohl dort für Fünfzehnjährige *Schulpflicht* herrscht, wurden für PISA nur 83 % des Jahrgangs

als eingeschrieben registriert (TR, S. 168). Etwa 2,5 % erklären sich durch die italienischen Berufsschulen, die von vornherein ausgeschlossen wurden. Einen größeren Teil der „verschwundenen 17 Prozent“ erklärt sich Putz damit,

> dass auch all jene Berufsschulen ausgeschlossen wurden, an denen zum Testzeitpunkt kein Unterricht stattgefunden habe. Das Pädagogische Institut habe aber lange vorher gewusst, an welchen Schulen am Testtag kein Unterricht stattgefunden habe. Damit konnte man das Ergebnis bewusst steuern. [Südtiroler Tageszeitung, 29./30. 1. 2005][15]

2.4 Ausschlüsse

Die internationalen Regeln erlaubten den Teilnehmerstaaten, bis zu 5 % der Grundgesamtheit vom Test auszuschließen, und zwar bis zu 0,5 % aus organisatorischen Gründen und bis zu 4,5 % wegen geistiger oder körperlicher Behinderung oder wegen ungenügender Beherrschung der Testsprache (TR, S. 47 f.). Über den Ausschluss wegen „intellectual disabilities“ hatten die Schulleiter zu entscheiden. Ein zusätzlicher Ausschlussgrund durfte in nationaler Verantwortung festgelegt werden (TR, S. 183 f.): die Tschechische Republik hat Schüler ausgeschlossen, die längere Zeit nicht am Unterricht teilgenommen hatten, Luxemburg frisch Zugewanderte, Dänemark, Finnland, Griechenland, Irland und Polen Schüler mit Lese-Rechtschreib-Schwäche, Dänemark zusätzlich Schüler mit *acalculia*.

Innerhalb der OECD sind diese Möglichkeiten sehr unterschiedlich ausgeschöpft worden: die Ausschlussquote streut von 0,7 % in der Türkei bis 7,3 % in Spanien und den USA (TR, S. 169). Auch Kanada, Dänemark und Neuseeland überschreiten die 5 %-Grenze. Diese Regelverstöße werden im Technischen Bericht vermerkt (TR, S. 241 ff.), haben aber keine weiteren Konsequenzen: die Daten aus den betroffenen Staaten werden uneingeschränkt in die Auswertung einbezogen.

2.5 Sonderschüler

Unabhängig von allen Ausschlusskriterien sieht das Testdesign vor, dass in Sonderschulen ein spezielles einstündiges Testheft eingesetzt werden kann. Diese Option wurde von nur sieben mitteleuropäischen Staaten genutzt. In allen übrigen Staaten sind Lernbehinderte angeblich in Regelschulen integriert – nachforschenswert wäre, inwieweit sie mit fünfzehn Jahren tatsächlich noch zur Schule

[15] Amüsant auch der Hinweis von Putz, welche Schlussfolgerungen man ziehen müsste, wenn PISA ernst zu nehmen und das Südtiroler Schulsystem tatsächlich weltweit das beste wäre: zum Beispiel die Lehrerausbildung abzuschaffen. In Südtirol konnte man sich, zumindest bis vor kurzem, im Anschluss an ein beliebiges Fachstudium durch einen *Aufsatzwettbewerb* für eine Lehrerstelle qualifizieren.

gehen. Ob sie vom Test ausgeschlossen werden, wird auf Schulebene entschieden; dass das zu Verzerrungen führen kann und Manipulationsmöglichkeiten eröffnet, liegt auf der Hand.

Prenzel begründet den Einsatz der Kurzhefte ethisch:

> Wir wollten nicht mit dem großen Testheft in die Sonderschulen, wir wollten es menschlicher machen [taz vom 9. 11. 2006].

Sofern man in Staaten ohne Sonderschulen ähnlich rücksichtsvoll war und Schüler mit Lernschwierigkeiten, in Ermangelung eines Kurzhefts, ganz vom Test ausgeschlossen hat, sind solche Schüler dort im internationalen Vergleich unterrepräsentiert.

Aber auch innerhalb der sieben Staaten, die das Kurzheft eingesetzt haben, sind die Stichproben kaum vergleichbar. In Österreich wurden 0,9 % aller Probanden mit dem Kurzheft getestet, in Ungarn dagegen 6,1 %. Die mittlere Leistung der Schüler, die das einstündige Heft bearbeitet haben, streut enorm, zum Beispiel im Lesen von 215 in Österreich bis 397 in Ungarn. Das muss nicht überraschen, denn in Österreich wurden 1,6 % der Grundgesamtheit, in Ungarn aber 3,9 % *ganz* vom Test ausgeschlossen: Wer in Österreich den Kurztest bearbeitet hat, wäre in Ungarn wahrscheinlich überhaupt nicht getestet worden; wer in Ungarn zum Kurztest eingeteilt wurde, wäre in Österreich wahrscheinlich zum regulären Test herangezogen worden.

Wie der Vergleich zwischen Deutschland (1,9 % Ausschlüsse, 3,6 % Kurztests, 287 Leseleistung im Kurztest) und den Niederlanden (1,9 %, 3,0 %, 380) zeigt, erklären die offiziell registrierten Ausschlussquoten aber nicht jeden Leistungsunterschied: entweder sind die niederländischen Sonderschulen den deutschen haushoch überlegen, oder es gibt weitere, weniger offensichtliche Ungleichmäßigkeiten in Zielgruppendefinition, Stichprobenziehung oder/und Testdurchführung. Prais (2003), der die Uneinheitlichkeiten bei der Durchführung von PISA 2000 und die dafür gegebenen oder rekonstruierbaren Erklärungen „kafkaesk“ (S. 149) nennt, weist darauf hin, dass *innerhalb* der Sonderschulen leistungsschwächere Schüler vom Test ausgeschlossen werden konnten (S. 158).

Im Gegensatz zu anderen Verzerrungen und Unsicherheiten, die sich aus nicht nachprüfbaren Ausschlüssen ergeben, kann die mit dem Einsatz von Kurzheften verbundene Ungenauigkeit ein Stück weit quantifiziert werden. Wenn Kurztestteilnehmer einheitlich von der Auswertung ausgenommen würden, stiege die Leseleistung in Deutschland um 7,6 Punkte; im OECD-Ranking würde Deutschland allein dadurch vom 18. auf den 12. Rang vorrücken.

Der Wert dieser Abschätzung liegt darin, dass sie beispielhaft zeigt, wie stark sich Uneinheitlichkeiten der Stichprobenziehung auf nationale Mittelwerte auswirken können. Bei anderen Ausschlusskriterien ist eine solche Abschätzung nicht möglich: Legastheniker zum Beispiel wurden in fünf Staaten nicht getestet, in den übrigen Staaten aber nicht als solche markiert, so dass keine

quantitative Abschätzung der Auswirkung auf das Leseergebnis möglich ist. Ziel dieser Abschätzung ist *nicht*, zu behaupten, der 12. Rang sei das korrekte Ergebnis; ich bestreite vielmehr die Seriosität einer wie auch immer konstruierten Rangordnung.[16]

Noch stärker als auf Mittelwerte wirken sich Uneinheitlichkeiten bei der Stichprobenziehung auf Breite und Form der Kompetenzverteilung aus. Als Beispiel sei der Anteil besonders schwacher Schüler betrachtet, willkürlich definiert als Schüler, die weniger als 400 Punkte erreicht haben, also um mehr als eine Standardabweichung unter dem internationalen Schnitt von 500 Punkten liegen. Mit Sonderschulen liegt dieser Anteil OECD-weit bei 16,2 %, in Deutschland bei 16,8 %. Ohne Sonderschulen sinkt er OECD-weit auf 15,8 %, in Deutschland auf 14,1 %. Ob der Anteil besonders schwacher Schüler in einem bestimmten Staat über oder unter dem internationalen Durchschnitt liegt, kann also entscheidend von uneinheitlich gehandhabten, unkontrollierbaren Details der Stichprobenziehung abhängen.

2.6 Verstöße gegen Teilnahmequoren

Auf beiden Stufen der Stichprobenziehung gibt es Ausfälle: Schulen lehnen die Teilnahme ab, und Schüler erscheinen nicht zum Test oder bleiben nicht bis zum Schluss. In PISA wird eine Teilnahmequote von 85 % aller Schulen gefordert. Quoten zwischen 65 % und 85 % können durch nachbenannte Ersatzschulen geheilt werden. Wenn auch mit Ersatzschulen keine „akzeptable“ Quote von mindestens 85 % erreicht wird, greift eine seltsame, nur graphisch mitgeteilte Regel (TR, S. 49), die bewusst offen lässt, wie ein solcher „intermediärer“ Fall zu handhaben ist, und das weitere Prozedere damit politischer Aushandlung (TR, S. 238 ff.) anheim stellt.

Schulen, in denen weniger als 50 % der ausgewählten Schüler den kognitiven Test abgeschlossen haben, werden als nicht teilnehmend gezählt; sofern diese Quote über 25 % lag, werden die Ergebnisse der teilnehmenden Schüler nichtsdestoweniger in den internationalen Datensatz aufgenommen und sogar hoch

[16] Das hatte ich in W1 leider nicht hinreichend deutlich gemacht; in der öffentlichen Rezeption ist die Verschiebung um 6 Rangplätze als *Korrektur* eines PISA-Ergebnisses missverstanden und als ein Schönrechnen der deutschen Ergebnisse kritisiert worden. Die bildungs- und forschungspolitische Sprecherin einer Bundestagsfraktion vermutete sofort (wenige Stunden nach Erscheinen des ersten Zeitungsberichts und sehr wahrscheinlich nur auf diesen gestützt) eine heimliche Agenda: mein Ziel sei, „geplante Reformen im Schulsystem anzugreifen, die auf die Integration von Benachteiligten zielten“ (Spiegel online, 8. 11. 2006). Dieses Argumentationsmuster ist erschreckend weit verbreitet: mit der Unterstellung, wer die Seriosität und Wissenschaftlichkeit einer Studie infrage stelle, bestreite auch die Notwendigkeit von Reformen, kann man es in Deutschland sogar in eine pädagogische Fachzeitschrift bringen (Bethge 1999).

gewichtet. Landesweit wird eine Schülerteilnahmequote von 80 % gefordert (TR, S. 48–50).

Das letztgenannte Quorum wurde 2003 von allen OECD-Staaten außer Großbritannien erreicht, von einigen aber nur knapp: in Australien, Österreich, Kanada, Irland, Polen und den USA lag die Schülerteilnahemequote unter 84 %, in acht weiteren Staaten unter 90 %. Die Schulteilnahmequote lag in der Mehrheit der Teilnehmerstaaten spätestens nach Heranziehung von Ersatzschulen zwischen 98 % und 100 %, in Belgien, Griechenland, Irland und Japan und Mexiko aber unter 96 %, in Australien, Frankreich, den Niederlanden und Norwegen unter 91 %. Großbritannien verfehlte auch dieses Kriterium (anfänglich 64,3 %, mit Ersatzschulen 77,4 %) und wurde letztlich aus allen OECD-Ranglisten ausgeschlossen (TR, S. 171–173).

Die niedrigen Schulteilnahmequoten der USA (64,9 %, mit Ersatz 68,1 %) und Kanadas (80,0 %, 84,4 %) wurden jedoch hingenommen, obwohl darüber hinaus auch unerlaubt hohe *within school exclusion rates* festgestellt wurden (TR, S. 238 ff.). Ähnlich war schon in PISA 2000 zugunsten der USA und Großbritanniens verfahren worden. Diese a posteriori und ad hoc beschlossene Missachtung selbstgegebener Regeln hat alsbald Kritik auf sich gezogen (von Collani 2001, S. 234 ff.; Prais 2003 S. 149 f.). Der PISA-Projektleiter warf Prais daraufhin „misunderstanding“ und „a lack of research“ vor. Er gestand zu, dass niedrige Antwortraten „a matter for concern“ und „a threat of bias“ seien, berief sich aber auf Zusatzuntersuchungen, die für Großbritannien keine Korrelation zwischen Teilnahmequoten und Leistungsfähigkeit fanden. Für PISA 2003 kündigte er eine ergänzende Überprüfung anhand landesweiter Lernkontrollen an (Adams 2003, S. 383 ff.). Sie ergab, dass in der PISA-Stichprobe sowohl besonders schwache als auch besonders starke Schulen unterrepräsentiert sein könnten, und dass die Nichtteilnahme einzelner Schüler wahrscheinlich eine Verzerrung der Leistungsmittelwerte bewirkt, die nicht durch Höhergewichtung anderer Schüler ausgeglichen werden kann. Dieses Ergebnis, das Prais in einem wichtigen Punkt recht gibt, wurde an entlegener Stelle im Technischen Bericht veröffentlicht (TR, S. 246 f.).

Für die USA wurde in PISA 2003 nichtsdestoweniger erneut mit der Nichtauffindbarkeit einiger Korrelationen argumentiert, aus der man entnehmen könne, dass der Datensatz „relativ wenig“ durch Schulnichtteilnahme verzerrt sei (TR, S. 247 f.). Immerhin konzediert Schulz (2006, Punkt 3), die Entscheidung, auch Daten aus Ländern einzubeziehen, die zuvor festgelegte Quoren nicht erfüllt haben, „might be debatable“ – wohl das expliziteste Eingeständnis eines Schwachpunktes, das man vom Konsortium erwarten darf.

2.7 Unzureichende Teilnahmequoren

Der Genauigkeitsanspruch von PISA liegt, grob gesprochen, im Prozentbereich. Daher ist zu vermuten, dass nicht allein die Verletzung anfangs festgelegter

Regeln problematisch ist, sondern dass diese Regeln selbst, indem sie Ausfälle von 20 bzw. 35 Prozent erlauben, unhaltbar großzügig sind.

Das Ausmaß der dadurch ermöglichten Verzerrungen ist schwer zu bestimmen, denn von den Schulen und Schülern, die die Teilnahme verweigert haben, liegen ja keine Daten vor. Im folgenden kann nur gezeigt werden, dass eine erhebliche Korrelation auf der Ebene ganzer Schulen besteht: je schwächer die Schule, desto geringer – im statistischen Mittel – die Teilnahmequote. Dieser Effekt wird in der offiziellen Auswertung zwar ausgeglichen: der Kehrwert der schulischen Teilnahmequote wird in das statistische Gewicht der Probanden hineingerechnet; Probanden von teilnahmeschwachen Schulen werden also in der weiteren Auswertung höher gewichtet. Bei diesem Ausgleich wird aber angenommen, dass die teilnehmenden Schüler repräsentativ für ihre Schule sind, dass also innerhalb einer Schule die Teilnahmeneigung nicht mit der Leistungsfähigkeit korreliert ist. Das ist unrealistisch. Die Teilnahmequoten kommen durch die individuelle Entscheidung einzelner Schüler zustande, am Test teilzunehmen oder nicht. Es gibt nicht den geringsten Grund, dass sich die in aggregierten Daten nachweisbare starke Korrelation zwischen Teilnahmeneigung und Leistungspotential nicht in die einzelnen Schulen hinein fortsetzt.

Auf Makroebene ist der Nachweis einer Korrelation von Teilnahmequote und Leistungsmittelwert leicht zu führen:

(1) Die Korrelation lässt sich an der Schulform festmachen. Zum Beispiel betrug die Teilnahmequote in Bayern im Mittel über alle Schulen 90 %, an Berufsschulen aber nur 68 % (Auskunft des Kultusministeriums laut G. Lind 2005). Für Frankreich lässt sie sich aus dem Datensatz erschließen: am Lycèe gènèral oder technologique 88 %, am Collège 77 %.

(2) Die Korrelation lässt sich auf der Ebene ganzer Schulen zeigen. Beispielsweise kann man aus dem Datensatz erschließen, dass der Sollumfang der Schulstichproben in Deutschland $n = 25$ betrug.[17] Schulen, an denen tatsächlich 25 Schüler am Test teilgenommen haben, erzielten im Mathematiktest 553 Punkte; Schulen mit 23 oder weniger Teilnehmern 505 Punkte oder weniger. Abbildung 1 zeigt weitere Daten für vier Staaten, in denen die Korrelation besonders deutlich ist. Dieses Argument steht allerdings unter einem kleinen Vorbehalt: da der Sollumfang nicht auf Schulebene dokumentiert ist, ist es möglich, dass in einzelnen Schulen weniger als n Fünfzehnjährige eingeschrieben sind.

(3) Hier helfen die französischen Daten, denn in Frankreich gibt es eigene Strata für „kleine“ und „sehr kleine“ Schulen. Das und weitere, indirekte Hinweise sprechen dafür, dass an den übrigen 144 Schulen eine Stichprobe vom einheitlichen Umfang $n = 32$ erreichbar war. Wenn man sich auf diese 144 Schu-

[17] Einer Insider-Information zufolge wurden in Deutschland 35 Schüler pro Schule herangezogen. Den Widerspruch kann ich einstweilen nicht auflösen.

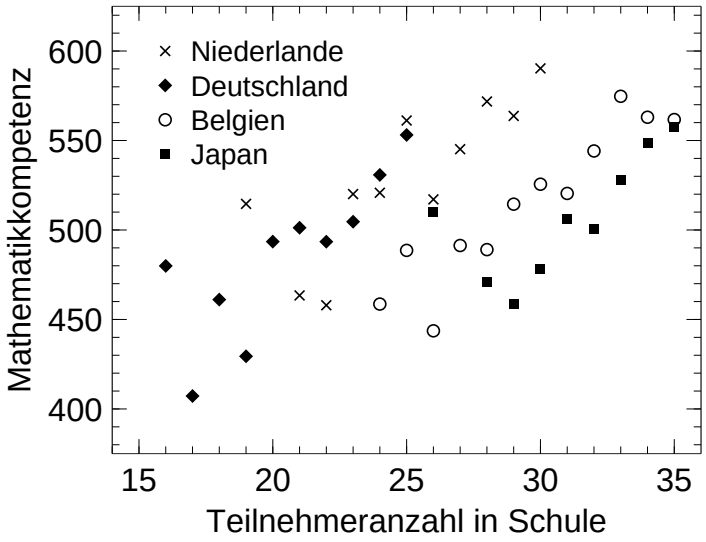

Abbildung 1: Mathematikkompetenz, jeweils gemittelt über alle Schulen eines Staates, die eine bestimmte Teilnehmerzahl erreicht haben. Gezeigt sind nur Datenpunkte, die mindestens drei Schulen repräsentieren. Man kann erschließen, dass die nationale Sollvorgabe für den Stichprobenumfang in Belgien und Japan 35, in den Niederlanden 30 und in Deutschland 25 betrug.

len beschränkt, findet man eine deutliche Korrelation zwischen Teilnehmerzahl und Leistung (in Mathematik $r = 0.41$).

Sowohl die Teilnahmeneigung als auch die Testleistung hängen stark von der Schulform ab. Die Leistungsmittelwerte von Lycèe und Collège unterscheiden sich um 121 Punkte. Dieser Unterschied hängt damit zusammen, dass in Frankreich regulär mit 15 Jahren der Übergang vom Collège in eine weiterführende Schule stattfindet. Fünfzehnjährige, die noch ein Collège besuchen, haben mit überproportionaler Wahrscheinlichkeit mindestens einmal eine Klasse wiederholt. Das zeigt, wie empfindlich PISA-Ergebnisse davon abhängen, dass die Stichprobe eine repräsentative Mischung verschiedener Schulformen darstellt, und dass innerhalb der einzelnen Schulen die Probanden im korrekten Verhältnis aus verschiedenen Jahrgangsstufen gezogen werden.

Wenn sich die Korrelation von Teilnahmeneigung und Testleistung in ähnlicher Größenordnung in die einzelnen Schulen hinein fortsetzt, kann man abschätzen, dass Leistungsmittelwerte in vielen Staaten um etliche Punkte überschätzt werden. Zugleich werden der Anteil sehr schwacher Schüler und die Varianz der Leistungsfähigkeit systematisch unterschätzt.

(4) Übrigens wird eine Verzerrung durch ungleichmäßige Teilnahmequoten auch von den PISA-Mitarbeitern Monseur und Wu für wahrscheinlich gehalten, weshalb sie in einem Konferenzbeitrag (2002, „incomplete and should not be quoted or cited“) vorgeschlagen haben, die kognitiven Fähigkeiten nicht teilnehmender Schüler aufgrund von Hintergrunddaten zu *schätzen*.

Tabelle 1: Mädchenanteil in der Bevölkerung und in der PISA–Stichprobe. Der Mädchenanteil in der Bevölkerung (U. S. Census Bureau o. J.) ist über die Altersklassen 10–14 und 15–19 Jahre gemittelt (die Abweichung beträgt nur in zwei Staaten mehr als 0,2 % und überall weniger als 0,5 %). Zum Mädchenanteil von PISA 2003 ist in Klammern der Standardfehler σ angegeben, der sich aus Stichprobengröße ergibt. Die Differenz zwischen Bevölkerungsjahrgang und PISA–Stichprobe ist in Vielfachen von σ angegeben. Vor Normierung auf σ wurde der Differenzbetrag um 0,5 % reduziert, um etwaige Ungenauigkeiten der Bevölkerungsdaten konservativ abzuschätzen.

Staat	Mädchenanteil		
	Bevölkerung	PISA 2003	Differenz
Südkorea	47,7 %	40,5 % (0,7 %)	$-10,0\,\sigma$
Türkei	49,2 %	45,0 % (0,7 %)	$-5,2\,\sigma$
Ungarn	48,9 %	47,2 % (0,7 %)	$-1,6\,\sigma$
Schweden	48,6 %	50,0 % (0,7 %)	$+1,1\,\sigma$
Finnland	48,9 %	50,1 % (0,7 %)	$+1,2\,\sigma$
Luxemburg	48,7 %	50,8 % (0,8 %)	$+2,0\,\sigma$
Dänemark	48,8 %	50,9 % (0,8 %)	$+2,1\,\sigma$
Japan	48,8 %	51,7 % (0,7 %)	$+3,3\,\sigma$
Spanien	48,6 %	50,8 % (0,5 %)	$+3,4\,\sigma$
Griechenland	48,6 %	51,7 % (0,7 %)	$+3,5\,\sigma$
Frankreich	48,9 %	52,6 % (0,8 %)	$+4,2\,\sigma$
Kanada	48,9 %	50,7 % (0,3 %)	$+4,3\,\sigma$
Portugal	48,0 %	52,4 % (0,7 %)	$+5,3\,\sigma$
Italien	48,6 %	51,9 % (0,5 %)	$+6,0\,\sigma$
Mexiko	49,4 %	51,8 % (0,3 %)	$+6,9\,\sigma$

2.8 Geschlechterverteilung

Die Repräsentativität der Stichprobe lässt sich auch anhand der Variablen Geschlecht und Geburtsdatum überprüfen. Dabei ist zu berücksichtigen, dass etwas mehr Jungen als Mädchen geboren werden. In der OECD liegt der Mädchenanteil in der Altersklasse der Zehn- bis Zwanzigjährigen zwischen 47,5 % in Südkorea und 49,3 % in Mexiko (U. S. Census Bureau 2006, vgl. NCHS 2006). Im PISA-Datensatz streut der Mädchenanteil hingegen zwischen 40,5 % in Südkorea und 52,6 % in Frankreich. Die konservative Abschätzung in Tabelle 1 zeigt, dass die Abweichung zwischen PISA-Stichprobe und Grundgesamtheit in fünf von dreißig OECD-Staaten mehr als 5σ beträgt. Dass diese Abweichungen allein auf stichprobenbedingte Variation zurückgehen, ist jenseits aller Plausibilität.

In Südkorea ist bereits der Mädchenanteil von nur 47,7 % im Altersjahrgang auffällig. Möglicherweise wird diese Anomalie in den gemittelten Daten des U. S. Census Bureau sogar noch unterschätzt; andere Quellen deuten darauf

hin, dass der Mädchenanteil im PISA-2003-Jahrgang zwischen 46 % und 47 % liegt.[18] Auch nach Berücksichtigung dieser Anomalie bleibt jedoch eine ganz erhebliche Diskrepanz von ca. 9σ zwischen dem Altersjahrgang und der PISA-Stichprobe.

Eine solche Abweichung kann drei verschiedene Ursachen haben: (1) geschlechtsabhängige Einschreibungsquoten, (2) Fehler bei der Stichprobenziehung und (3) geschlechtsabhängige Teilnahmequoten. Laut Technischem Bericht (TR, S. 171 ff.) gehen in Südkorea 99,94 % aller Fünfzehnjährigen zur Schule, die Schulteilnahmequote betrug 100 %, und die Schülerteilnahmequote hatte den Spitzenwert von 98,81 %. Wenn diese Angaben zutreffen, können die Ursachen (1) und (3) ausgeschlossen werden; dann muss es zu eklatanten Fehlern bei der Stichprobenziehung gekommen sein. Auffällig ist überdies die Altersverteilung: 29,7 % der Schüler sind im ersten Drittel, 38,7 % im letzten Drittel des getesteten Jahrgangs geboren. Bei funktionierender Plausibilitätskontrolle hätten die koreanischen Daten nicht in die Auswertung einbezogen werden dürfen.

Schulz (2006, Punkt 4) berichtet, die Anomalie in der koreanischen Stichprobe sei sofort bemerkt, aber nach Vergleich mit nationalen Statistiken für plausibel befunden worden. Allerdings bezieht sich Schulz auf einen älteren Textstand; erst später habe ich die hier zusammengefasste Literaturrecherche unternommen, die konsistent dagegen spricht, dass die Anomalie der PISA-Stichprobe vollständig durch eine anomale Zusammensetzung der Bevölkerung erklärt werden kann.[19]

Der Mädchenanteil von nur 45,0 % in der Türkei dürfte hingegen einen realen Geschlechterunterschied in der Einschreibungsquote wiederspiegeln und bestätigt, dass die Wahl der PISA-Grundgesamtheit ungeeignet ist, das Gesamtergebnis des türkischen Erziehungssystems zu untersuchen. Schwerer zu beurteilen ist, inwieweit der zu hohe Mädchenanteil der PISA-Stichprobe in etlichen entwickelteren Staaten auf unterschiedlicher Schuleinschreibung und inwieweit auf

[18] Weltweit liegt der Mädchenanteil bei der Geburt bei ungefähr 48,8 %. Das Problem der „fehlenden Mädchen“ in Asien ist Gegenstand intensiver Forschung (CEPED 2006). Zum Teil kann es möglicherweise als Auswirkung von Hepatis-B erklärt werden (Oster 2005); in Südkorea beruht es jedoch überwiegend auf selektiver Abtreibung namentlich ab der zweiten Schwangerschaft (Song 1998, Kim 2004). Diese Praxis hat Mitte der 1990er Jahre ihren Höhepunkt erreicht und zu einem Mädchenanteil von knapp unter 46,5 % geführt (CEPED 2006).

[19] Neuwirth (Mail vom 22. 12. 2006) hat mich darauf hingewiesen, dass in der PISA-Stichprobenziehung das Geschlecht nicht als Stratifikationsvariable herangezogen wird. Das sei ungeschickt, zumal für Staaten wie Südkorea, wo es einen hohen Anteil reiner Jungen- und Mädchenschulen gibt. Rechnerisch rücke das den Mädchenanteil in der 2003er Stichprobe in die Nähe des statistisch nicht ganz Unwahrscheinlichen. Gegen eine rein stochastische Erklärung spricht aber, dass auch die Altersverteilung anomal ist und dass ähnliche Anomalien schon in der 2000er Stichprobe auftraten.

unterschiedlicher Testteilnahme beruht. Bei beiden Ursachen ist eine Korrelation mit der Leistung zu vermuten.

2.9 Umgang mit unvollständigen Testheften

PISA 2003 bestand aus einem zweistündigen „kognitiven“ Test, gefolgt von einer knappen Stunde, in der mit Fragebögen verschiedenste Hintergrundvariable erhoben wurden. In beiden Teilen des Tests ist damit zu rechnen, dass Schüler wegen Unpässlichkeit oder Unlust ihre Teilnahme abbrechen.[20]

Wie wirksam Schule und Testleitung dem entgegenwirken können, dürfte erheblich von national unterschiedlichen kulturellen, organisatorischen und rechtlichen Randbedingungen abhängen. Tatsächlich variiert die Quote der aus dem Datensatz erkennbaren Testabbrüche zwischen exakt 0 % (Finnland, Luxemburg, Polen) und 0,9 % (Großbritannien).[21] Verglichen mit der Quote von Schülern, die gar nicht erst zum Test antreten, scheinen Testabbrüche ein untergeordnetes Problem zu sein; fraglich ist jedoch, ob sie international einheitlich erfasst und ausgewertet wurden.

Die Schweizer Projektleitung hat mir nämlich mitgeteilt, „dass überall das Testheft erst nach den 2 Stunden Testsitzung eingesammelt wird, somit die Schüler überhaupt nicht vorzeitig abgeben können“ (C. Zahner-Rossier, Mail vom 28. 3. 2006). Diese Auskunft ist nicht nur weltfremd, sondern widerspricht den internationalen Vorgaben: jeder Testleiter sollte für jeden Schüler und für jede der drei Teststunden festhalten, ob der Schüler anwesend, abwesend, oder partiell (zwischen 5 und 55 Minuten) anwesend war (OECD 2003b, S. 15); aufgrund dieser Angaben wurden bei Schülern, die vorzeitig abgegeben haben, nicht erreichte Aufgaben als nicht gestellt gewertet.[22] Wie soll man glauben, dass solche Regeln einheitlich umgesetzt wurden, wenn eine nationale Projektleitung sie nicht kennt und sogar jeden Bedarf für eine Regelung bestreitet?

Laut Regelwerk war als Teilnehmer zu werten, wer an der kognitiven Testung teilgenommen hatte. In den internationalen Datensatz sollten demnach auch Schüler aufgenommen werden, die den Test vor oder während der Fragebogen-Sitzung abgebrochen haben (TR, S. 50, 162). In Kanada haben 9,7 % aller Schüler das *Student Questionnaire* überhaupt nicht bearbeitet, in Deutschland 2,1 %

[20] Bericht eines Hamburger Schülervaters über die Durchführung von PISA: Die Testung fand im Anschluss an regulären Unterricht statt; die Schüler durften nach Hause gehen, wenn sie „fertig“ waren; die ersten Schüler gaben nach wenigen Minuten ab.

[21] Drei oder mehr Aufgaben unmittelbar vor dem Testende als „not applicable“ markiert.

[22] Die Angaben im Auswertehandbuch (DAM, S. 243/248) sind irreführend. Sie lassen vermuten, eine Untermenge der „non-reached“-Codes sei als „not-applicable“ zu verstehen. R. Adams (Mail vom 17. 1. 2007) teilt jedoch mit, dass nicht erreichte Aufgaben für Schüler, die vorzeitig abgegeben haben, im veröffentlichten Datensatz als „not-applicable“ kodiert sind.

(nach Ausschluss der Sonderschüler, sonst 5,6 %), überall sonst deutlich weniger. Schulz (2006, Punkt 5) erklärt die kanadische Anomalie damit, dass einige Provinzregierungen rechtliche Einwände gegen den Fragebogen hatten. Für die Ausfälle in Deutschland wurde noch keine Erklärung gegeben. Der Vergleich mit PISA 2000, wo noch eine ganze Reihe von Ländern ähnlich hohe Ausfallquoten hatte, könnte darauf hindeuten, dass Deutschland eine Verengung des Teilnahmekriteriums nicht umgesetzt hat.

Fast überall liegen die Testleistungen derjenigen Schüler, die das Questionnaire nicht bearbeitet haben, unter dem nationalen Durchschnitt, oft um 100 oder mehr Punkte. Ohne diese Schüler steigt die mittlere Leistung in Kanada um 3,6 Punkte, in Deutschland (ohne Sonderschulen) um 1,5 Punkte, in den übrigen OECD-Staaten im Mittel um nur 0,2 Punkte.

Aus verschiedensten Gründen lassen manche Schüler einzelne Fragen unbeantwortet. In Polen sind jedoch sieben Fragen von keinem einzigen Schüler *nicht* beantwortet worden, und es findet sich kein einziger polnischer Schüler, der weniger als 25 Fragen gültig beantwortet hat. Solange keine andere Erklärung für diese Anomalie glaubhaft gemacht wird, muss als maximale Verzerrung unterstellt werden, dass Polen Schüler, die die Fragebögen nicht oder sehr unvollständig bearbeitet haben, nicht zum internationalen Datensatz beigetragen und dadurch das nationale Leistungsmittel geschönt hat.

3 Wo kommen die Punkte her?

In diesem Teil wird eine selbstkonsistente, ohne Rückgriff auf Spezialliteratur lesbare Beschreibung des Skalierungsverfahrens gegeben, mit dem in PISA von kognitiven Testergebnissen auf Kennzahlen geschlossen wird, die als Aufgabenschwierigkeiten und Schülerkompetenzen gedeutet werden. Nach einem Überblick über Testdesign und Datenstruktur (3.1 ff.) werden Antwort- und Bevölkerungsmodelle eingeführt (3.3 ff.). Die Parameterschätzung beruht auf dem Satz von Bayes (3.5 f.); die Berechnung von Populationsmittelwerten auf einer Monte-Carlo-Integration, deren Stützpunkte als „plausible Werte“ veröffentlicht werden (3.9). Um Standardfehler zu senken, wird die Schätzung mit Hintergrundvariablen konditioniert (3.7) und auch auf Probanden angewandt, die das Testgebiet gar nicht zu bearbeiten hatten (3.10). Die Komplexität dieses Verfahrens täuscht leicht darüber hinweg, dass die Kompetenzwerte im Kern ganz einfach den Anteil richtig gelöster Aufgaben messen – jede Klassenarbeitskorrektur ist differenzierter (3.12). Die Umrechnung zwischen dem Anteil richtiger Aufgabenlösungen und der nach außen kommunizierten Punkteskala zeigt, wie empfindlich PISA-Ergebnisse von der Vergleichbarkeit sämtlicher Randbedingungen abhängen (3.14).

Diese gesamte Rekonstruktion steht unter dem Vorbehalt, dass sie auf einer mangelhaften Dokumentation beruht, die kein eigenständiges Nachvollziehen sämtlicher Rechnungen erlaubt (3.2). Das Konsortium ist herzlich eingeladen, verbliebene Missverständnisse zu korrigieren.

3.1 Testdesign, Datenerfassung und -aufbereitung

Der kognitive Test ist in Aufgabenstämme (*Units*) gegliedert. Ein Aufgabenstamm umfasst ein bis sieben einzelne Aufgaben (*Items*). Jeder Schüler bearbeitet rund 50 Aufgaben, die sich auf vier Testgebiete und innerhalb des schwerpunktmäßig untersuchten Gebiets Mathematik auf vier Untergebiete verteilen. Um eine gewisse Breite an Inhalten und Anforderungen abzudecken und um zu vermeiden, dass die Testergebnisse in extremer Weise vom „Funktionieren“ einzelner Aufgaben abhängen, werden dreizehn verschiedene Testhefte eingesetzt.[23] Um Ergebnisse aus verschiedenen Heften miteinander vergleichen zu können, wird jede der 165 Aufgaben in vier verschiedenen Heften gestellt, jeweils in einem anderen halbstündigen Block (Tab. 2).

[23] Als nützliche Nebenwirkung erschwert die Vielzahl der Testhefte wahrscheinlich das Abschreiben. Dennoch muss es, einer Kieler Dissertation zufolge, recht munter zugegangen sein: „da die Testsitzungen gemischtgeschlechtlich stattfanden“, wurde „die Geschlechtsidentität stärker aktiviert“ (Burba 2006, S. 154).

Tabelle 2: Testdesign von PISA 2003: Jedem Schüler wird eines von dreizehn Testheften zugewiesen. Jedes Heft enthält vier Blöcke. Jeder Block enthält Aufgaben aus einem der vier Gebiete Lesen (L), Mathematik (M), Naturwissenschaften (N) und problemlösendes Denken (P).

Zeitablauf	Testheft												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Stunde	M1	M2	M3	M4	M5	M6	M7	N1	N2	L1	L2	P1	P2
	M2	M3	M4	M5	M6	M7	N1	N2	L1	L2	P1	P2	M1
Pause													
2. Stunde	M4	M5	M6	M7	N1	N2	L1	L2	P1	P2	M1	M2	M3
	L1	L2	P1	P2	M1	M2	M3	M4	M5	M6	M7	N1	N2
Pause													
3. Stunde	Fragebögen												

Dieses Testdesign bringt mit sich, dass Ergebnisse aus verschiedenen Heften oder Blöcken nicht unmittelbar miteinander verglichen werden können; Abbildung 2 zeigt, wie die unterschiedliche Mischung leichter und schwerer Aufgaben zu ganz unterschiedlich verteilten Lösungshäufigkeiten führen kann. Um dennoch zu einer eindimensionalen Bewertung der Schülerleistungen zu kommen, haben die Veranstalter von vornherein eine ganz bestimmte, modellabhängige Auswertung geplant. In PISA 2000 war diese Einengung extrem. Die wesentlich symmetrischere Anordnung der Aufgabenblöcke von PISA 2003 erleichtert Auswertungen, die sich unmittelbar auf die prozentualen Lösungshäufigkeiten einzelner Aufgaben stützen.

Als erster Schritt der Datenauswertung werden die Schülerhefte in nationaler Verantwortung von eigens geschulten Hilfskräften kodiert. Der Technische Bericht enthält zwei volle Kapitel (TR, S. 135 ff., 217 ff.) über „Coder Reliability Studies“. Mit beträchtlichem statistischem Aufwand wird dort gezeigt, dass es systematische Verzerrungen im Prozentbereich gibt; bei einzelnen Aufgaben in einzelnen Staaten erreicht die Verzerrung zig Prozent (TR, S. 232).

Die kodierten Schülerantworten aus kognitivem Test und Fragebögen werden in eine Datei zusammengeführt und an die internationale Projektleitung bei ACER (Australian Council of Educational Research) übermittelt. Dort werden die nationalen Ergebnisse zum internationalen Datensatz zusammengesetzt. Anschließend werden aus der Gesamtheit der Schülerantworten die Aufgabenschwierigkeiten und Schülerkompetenzen bestimmt. Dieser Schritt wird als die *Skalierung* des internationalen Datensatzes bezeichnet. Wegen der probabilistischen Natur der dabei verwendeten Modelle erhält man für die Schülerkompetenzen keine eindeutigen Zahlenwerte, sondern Wahrscheinlichkeitsdichten.

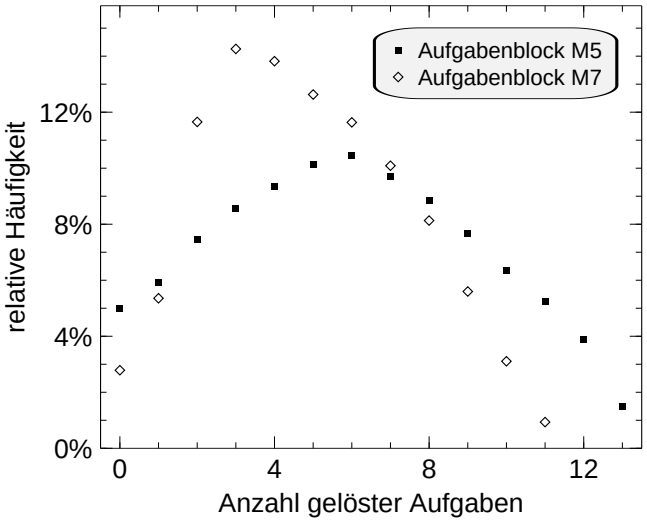

Abbildung 2: Diese Auftragung zeigt beispielhaft für zwei Aufgabenblöcke, wieviel Prozent der Schüler, im Mittel über 30 OECD-Staaten, wieviel Aufgaben richtig gelöst haben.

Für die gesamte weitere Auswertung werden diese Wahrscheinlichkeitsdichten durch fünf Zufallszahlen pro Schüler und pro Testgebiet, sogenannte *plausible values*, repräsentiert. Diese Zahlenwerte werden dem internationalen Datensatz hinzugefügt, der dann an die nationalen Projektzentren zurückübermittelt und nach Veröffentlichung der offiziellen Berichte auf der Website von ACER frei verfügbar gemacht wird.

3.2 Dokumentationsmängel im Technischen Bericht

Die nachfolgende Rekonstruktion des in PISA angewandten Skalierungsverfahrens steht unter dem Vorbehalt, dass sie sich auf eine mangelhafte Dokumentation stützt, namentlich auf Kapitel 9 des Technischen Berichts. Dieses Kapitel ist weitgehend unverändert aus dem Bericht zu PISA 2000 (Adams/Wu 2002) übernommen. Damals waren noch die Autoren der einzelnen Kapitel angegeben; für Kapitel 9 zeichnete allein Ray Adams, der Projektleiter des internationalen Konsortiums, verantwortlich.

Der Text beschreibt ein Verfahren, rechtfertigt es aber nicht. Es fehlt jede Erklärung theoretischer Grundlagen. Es fehlt jeder Hinweis auf Literatur, die dieses Defizit ausgleichen könnte. In der Hauptsache zitiert Adams überhaupt nur drei Arbeiten: Adams *et al.* (1997a), Adams *et al.* (1997b) sowie das Handbuch zum Computerprogramm ConQuest. Auch diese Texte gehen sofort in die technischen Details und unternehmen nicht die geringste Anstrengung, den konzeptionellen Rahmen zu erläutern. Im Literaturverzeichnis des Technischen Berichts ist kein einziges Überblickswerk zur Testtheorie angegeben.

Aber auch als Beschreibung eines Rechenverfahrens ist Kapitel 9 des Technischen Berichts weitgehend unbrauchbar. Simple Mathematik wird unter einer

ungeschickten, inkonsequenten und aufgeblähten Notation verschüttet. Die lineare Algebra wird unnötig strapaziert: Tupel ohne geometrische Bedeutung werden als Vektoren behandelt, nur um Produktsummen abkürzend notieren zu können, mit der Folge, dass sich der Leser mühsam zusammenreimen muss, über welche Indizes in diesen Skalarprodukten summiert wird. Um an einer einzigen Textstelle (TR, S. 121) zwei Gleichungen (9.3 und 9.4) kompakt zu schreiben, werden dreifach indizierte Kopplungskoeffizienten (b_{ijd}) unter vierfacher Vektortransposition als Matrix (B) verpackt.

Diese Matrix täuscht vor, jede Aufgabe trüge zu mehreren latenten Persönlichkeitseigenschaften (*traits*) bei. Indirekt, zum Beispiel durch Rückgriff auf Adams *et al.* (1997b), lässt sich jedoch erschließen, dass sie pro Zeile außer *einer* Eins nur Nullen enthält: Jede PISA-Aufgabe trägt zur Kompetenzschätzung in genau einem Testgebiet bei; sonst sind auch die Anhänge 12 ff. im Technischen Bericht nicht zu verstehen. Dass die multidimensionale Notation nichts als Popanz ist, bestätigt sich im weiteren Verlauf des Kapitels, wo die latenten Eigenschaften θ und die Aufgabenparameter ξ, ohne dass darauf hingewiesen würde, nicht mehr als Vektoren, sondern als Skalare auftreten.

Ohne Zusatzinformationen könnte man auch annehmen, die von Null verschiedenen Koeffizienten von B seien frei anpassbare Parameter. Das wäre durchaus sinnvoll, denn als variable Trennschärfen würden solche Parameter die Modellierung des Schülerverhaltens erheblich verbessern (4.2). Mindestens eine nationale Projektleitung hat ACER in diesem Sinne missverstanden: McCluskey und Zahner (2004, S. 14, Fußnote 13) setzen die Item-Response-Theorie fälschlich mit einem bestimmten Antwortmodell gleich und nehmen an, neben der Schwierigkeit sei auch die Trennschärfe ein Aufgabenmerkmal. In Wahrheit werden in PISA alle Aufgaben eines Gebiets mit ein und derselben Trennschärfe beschrieben, egal wie schlecht das zu den Daten passt.

Die einleitenden Seiten von Kapitel 9 des Technischen Berichts sind zu weiten Teilen wörtlich aus Kapitel 12 des ConQuest-Handbuchs extrahiert. Das erklärt, wie Adams dazu gekommen ist, die Skalierung unvollständig und in inadäquater Notation zu dokumentieren: Das Handbuch versucht, in größtmöglicher Allgemeinheit zu beschreiben, was ConQuest alles rechnen *kann*. Von der PISA-Dokumentation wäre aber Auskunft zu erwarten, was in der Auswertung tatsächlich gerechnet worden *ist*. Dieser Aufgabe hat sich Adams gar nicht erst gestellt.

Kapitel 9 im Technischen Bericht und Kapitel 12 im ConQuest-Handbuch erhellen sich in einigen Punkten gegenseitig; im direkten Vergleich wird erst so richtig deutlich, wie schlampig bei ACER gearbeitet wird. Konkretes Beispiel: Das Handbuch (S. 130f.) springt scheinbar unmotiviert zwischen gestrichenen und ungestrichenen Größen hin und her. Im PISA-Bericht wird die Notation $\mathbf{A}'$ durch A^T ersetzt und damit geklärt, dass die vormals gestrichenen Größen für transponierte Vektoren stehen. Nur: es sind nicht alle Striche durch T's er-

setzt worden. Beispielsweise ist in den Formeln (9.2) und (9.3) des Technischen Berichts versehentlich die gestrichene Variante beibehalten worden – im neuen Kontext völlig unverstehbar. Ungünstig ist auch, dass Vektoren nicht mehr durch Fettdruck kenntlich gemacht werden. Immerhin wird ein Index korrigiert, der im ConQuest-Handbuch seit acht Jahren falsch nachgedruckt wird (dort S. 131, Formel (2)).

Weitere Notationsmängel erschweren das Nachvollziehen. Allein auf den zwei Seiten 133–134 des ConQuest-Handbuchs fallen die folgenden auf: Das Durcheinander der in (11) und (12) alternativ zueinander eingeführten Hintergrundvariablen Y_n und W_n in (15)ff. Die nicht eingeführten überdachten Größen in (17) und (18). Die Verwendung des Buchstaben W in völlig verschiedenen Bedeutungen in (12) ff. und (21) ff. Für Funktionen fällt den Autoren kaum ein anderer Buchstabe als f ein: eine in den Technischen Bericht übernommene Formel (dort 9.10) verknüpft beispielsweise drei paarweise verschiedene Funktionen, die f_θ, f_x und f_x (sic!) heißen, wobei die Buchstaben θ und x in derselben Formel auch noch als Funktionsargumente auftreten und eine der Funktionen f_x zuvor ohne Subskript eingeführt worden ist. Trotz allen notationellen Aufwands bleibt völlig unberücksichtigt, dass nicht alle Probanden dieselben Aufgaben bearbeiten. Das ganze hat eher die Qualität eines im Verlauf einer mündlichen Diskussion improvisierten Tafelanschriebs, als dass es gängigen fachlichen Standards für mathematische Prosa genügt.

Die einzelnen Kapitel des Berichts sind mangelhaft aufeinander abgestimmt. Der abschließende Schritt der Skalierung, die Umrechnung auf externe Einheiten, bleibt im Skalierungskapitel unerwähnt und fällt stattdessen am Ende eines Ergebniskapitels ohne weitere Erklärungen vom Himmel. In der Einleitung eines späteren Kapitels wird ein falscher Rückblick auf die angeblich simultane Parameterschätzung gegeben. Wie planlos der Technische Bericht entstanden ist, zeigt sich auch im Vergleich von S. 129 und S. 206: eine volle Textseite wird wörtlich wiederholt, ohne dass ein Querverweis einen logischen Zusammenhang herstellt.

Jenseits all dieser spezifischen Mängel, die mit normaler Sorgfalt (zumal drei Jahre nach dem ersten Durchgang!) leicht vermeidbar gewesen wären, ist die Dokumentation ihrer ganzen Anlage nach nicht genau genug, um eine unabhängige Verifikation der numerischen Ergebnisse zu ermöglichen. Schon der *Input* der Skalierung lässt sich nicht reproduzieren (3.13). Daher ist weder auszuschließen, dass die folgende Rekonstruktion Missverständnisse enthält, noch, dass dem Konsortium bei der Implementierung Fehler unterlaufen sind. Köller (2006a) behauptet zwar, das deutsche Konsortium habe die ConQuest-Ergebnisse mit anderen Programmen validiert und habe „immer die Ergebnisse von ACER replizieren können“; die Schätzungen seien „identisch“. Das ist aber völlig unglaubwürdig, denn da die Skalierung auf einer *zufällig* gezogenen Unterstichprobe von 500 Probanden pro Staat beruht, ist es beliebig unwahrscheinlich,

dass unabhängig voneinander durchgeführte Skalierungen *identische* Schätzungen liefern.

3.3 Datenstruktur

In einem Test begegnen sich eine Menge $\mathcal{V}$ von Probanden („Versuchspersonen“) und eine Menge $\mathcal{I}$ von Aufgaben („Items“). Aber nicht jeder Proband bekommt jede Aufgabe vorgelegt. Jedem Probanden $v \in \mathcal{V}$ ist deshalb die Teilmenge $\mathcal{I}_v \subset \mathcal{I}$ derjenigen Aufgaben zugeordnet, die er tatsächlich zu bearbeiten hatte. Auf diese Weise kann sowohl die Verwendung verschiedener Testhefte, als auch die Löschung einzelner Aufgaben in einzelnen Staaten beschrieben werden.

Man könnte meinen, dieser erste Schritt zur Mathematisierung des Testgeschehens sei noch ziemlich elementar und unproblematisch, doch schon hier treten Interpretationsspielräume zutage: Wann genau gilt eine Aufgabe als gestellt? In PISA wird $\mathcal{I}_v$ je nach Auswertungsphase unterschiedlich abgegrenzt: als „nicht erreicht“ kodierte Aufgaben am Ende des kognitiven Tests werden bei der Ermittlung der Aufgabenschwierigkeiten ausgeschlossen, bei der Bestimmung der Schülerkompetenzen aber eingeschlossen und als falsch gewertet.[24]

Durch die Kodierung der Testantworten wird jedem Probanden $v \in \mathcal{V}$ für jede zu bearbeitende Aufgabe $i \in \mathcal{I}_v$ eine Antwortkategorie $\kappa_{vi} \in \mathcal{K}_i$ zugeordnet. In PISA 2003 sind 142 von 165 Aufgaben „dichotom“, das heißt, die Menge der möglichen Antwortkategorien ist $\mathcal{K}_i = \{\text{falsch, richtig}\}$. Bei 21 Aufgaben umfasst $\mathcal{K}_i$ außerdem die Kategorie „teilweise richtig“, bei zwei Aufgaben gibt es zwei Abstufungen teilrichtiger Antworten.

Wenn wir mit $\underline{\mathcal{I}}$ die Gesamtheit aller Aufgabensätze $\mathcal{I}_v$ und mit $\underline{\underline{\kappa}}$ die Gesamtheit aller Probandenantworten κ_{vi} bezeichnen, dann enthält $\langle \mathcal{V}, \underline{\mathcal{I}}, \underline{\underline{\kappa}} \rangle$ die vollständigen *Rohdaten* der kognitiven Testung.[25] Wenn von der Stichprobe auf die Grundgesamtheit geschlossen werden soll, sind außerdem die im Rahmen der zweistufigen Stichprobenziehung bestimmten Probandengewichte $w_v \in \mathtt{R}$ zu berücksichtigen.

---

[24] Mail von Adams vom 17. 1. 2007. Die Angaben im Auswertehandbuch (DAM, S. 248) und Technischen Bericht (TR, S. 161) sind unzureichend. Sie sagen nun, *dass* ausgelassene und nicht erreichte Aufgaben unterschiedlich berücksichtigt werden, aber nicht *wie*.

[25] Dieses Tupel ist ein bipartiter gefärbter Graph. Dieselbe mathematische Struktur findet man im elementaren Stundenplanproblem (z. B. Willemen 2002). Von dort entlehne ich die kompakte und flexible mengentheoretische Notation. In der Psychometrie ist es üblich, Probanden, Aufgaben, Antwortkategorien usw. durch natürliche Zahlen zu repräsentieren und einfach indizierte Größen als Vektoren, doppelt indizierte Größen wie $\underline{\underline{\kappa}}$ als Matrizen aufzufassen. Das ist unelegant und wird unübersichtlich, sobald nicht jeder Proband jede Aufgabe vorgelegt bekommt. In der hier gewählten Notation muss κ_{vi} gar nicht für alle Paare $v \in \mathcal{V}$, $i \in \mathcal{I}$ definiert sein; $\underline{\underline{\kappa}}$ ist eine *Funktion* $\{(i,v) | v \in \mathcal{V}, i \in \mathcal{I}_v\} \to \bigcup_{i \in \mathcal{I}} \mathcal{K}_i$.

PISA 2003 erstreckt sich über die Gebiete $\mathcal{G} = \{$Mathematik, Lesen, Naturwissenschaften, Problemlösen$\}$. Die weitere Unterteilung in vier Mathematik-Teilgebiete kommt erst *nach* der Skalierung zum Tragen (TR, S. 191) und kann hier unberücksichtigt bleiben. Ziel der Skalierung ist es letztlich, jedem Probanden $v \in \mathcal{V}$ für jedes Gebiet $g \in \mathcal{G}$ einen Kompetenzwert θ_{vg} zuzuordnen. Die multidimensionale Maskerade im Technischen Bericht deutet, wie in 3.2 beschrieben, die Möglichkeit an, dass einzelne Aufgaben zu mehr als einem Testgebiet beitragen können. Das ist nicht der Fall: jede Aufgabe i ist genau einem Gebiet g_i zugeordnet. Entgegen anderslautenden Textstellen (TR, S. 191: „multidimensional scaling“) werden PISA-Ergebnisse *nicht* mit einem mehrdimensionalen Modell skaliert, sondern es werden vier disjunkte Teiltests unabhängig voneinander mit je einem eindimensionalen Modell skaliert. Im folgenden genügt es deshalb, einen einzelnen Teiltest zu betrachten, konkret zumeist $g =$ Mathematik, so dass Abhängigkeiten von g nicht explizit notiert werden müssen.

3.4 Antwortmodelle

Ein Antwortmodell $\langle \mathcal{K}_i, \mathcal{P}_i, A_i \rangle$ enthält außer der schon eingeführten Menge $\mathcal{K}_i$ der möglichen Antwortkategorien einen Parameterraum $\mathcal{P}_i$ sowie eine Funktion A_i, die die Wahrscheinlichkeit angibt, dass die Antwort κ_{vi} von Proband v auf Aufgabe i in die Kategorie k fällt,

$$P(\kappa_{vi}{=}k) = A_i(k, \underline{\pi}_i, \theta_v)\,. \tag{1}$$

Jede Aufgabe hat einen Parametersatz $\underline{\pi}_i \in \mathcal{P}_i$, und jeder Proband einen Kompetenzwert $\theta_v \in \mathtt{R}$.

Zur konkreten Ausgestaltung werden in PISA je nach Anzahl der Antwortkategorien drei verschiedene psychologische Modelle verwendet:

$$A_i = \begin{cases} A_{\text{Rasch}} & \text{wenn } |\mathcal{K}_i| = 2, \\ A_{\text{PC3}} & \text{wenn } |\mathcal{K}_i| = 3, \\ A_{\text{PC4}} & \text{wenn } |\mathcal{K}_i| = 4\,. \end{cases} \tag{2}$$

Beim Rasch-Modell für dichotome Aufgaben ist $\mathcal{P}_i = \mathtt{R}$; es gibt genau einen Parameter $\underline{\pi}_i = \xi_i$, der die Schwierigkeit der Aufgabe beschreibt. Die Antwortwahrscheinlichkeiten sind

$$A_{\text{Rasch}}(\text{falsch}, \xi, \theta) = \frac{1}{1 + \mathrm{e}^{\theta - \xi}}\,, \tag{3}$$

$$A_{\text{Rasch}}(\text{richtig}, \xi, \theta) = \frac{\mathrm{e}^{\theta - \xi}}{1 + \mathrm{e}^{\theta - \xi}}\,. \tag{4}$$

Für die übrigen Aufgaben werden „Partial Credit"-Modelle verwendet. Für Aufgaben mit drei Antwortkategorien ist $\mathcal{P}_i = \mathbb{R}^2$, $\underline{\pi}_i = (\xi_{i1}, \xi_{i2})$, und die Lösungswahrscheinlichkeiten lauten

$$A_{\mathrm{PC3}}(\text{falsch}, (\xi_1, \xi_2), \theta) = \frac{1}{1 + \mathrm{e}^{\theta-\xi_1} + \mathrm{e}^{2\theta-\xi_1-\xi_2}}\,, \tag{5}$$

$$A_{\mathrm{PC3}}(\text{teilrichtig}, (\xi_1, \xi_2), \theta) = \frac{\mathrm{e}^{\theta-\xi_1}}{1 + \mathrm{e}^{\theta-\xi_1} + \mathrm{e}^{2\theta-\xi_1-\xi_2}}\,, \tag{6}$$

$$A_{\mathrm{PC3}}(\text{richtig}, (\xi_1, \xi_2), \theta) = \frac{\mathrm{e}^{2\theta-\xi_1-\xi_2}}{1 + \mathrm{e}^{\theta-\xi_1} + \mathrm{e}^{2\theta-\xi_1-\xi_2}}\,. \tag{7}$$

Durch analoges Hinzufügen eines weiteren Schwierigkeitsparameters erhält man das Modell PC4 für tetrachotome Aufgaben.

Ein erheblicher Teil der linearen Algebra im Technischen Bericht dient allein dem Zweck, diese drei Modelle durch *eine* einheitliche Formel darzustellen. Die Koeffizienten der dazu eingeführten Matrizen muss man sich aus Adams *et al.* (1997b) zusammensuchen; eine Bestätigung für die Rekonstruktion (5)ff. findet sich im deutschen Bericht (Prenzel *et al.* 2004b, im weiteren zitiert als D03b, S. 397).

Ein leitender Wissenschaftler des *Educational Testing Service* hält die Verwendung solcher Modelle für einen das Geschäftsmodell der Testindustrie bedrohenden Anachronismus:

> It is only a slight exaggeration to describe the test theory that dominates educational measurement today as the application of 20th century statistics to 19th century psychology. Sophisticated estimation procedures [...] applied within psychological models that explain problem-solving ability in terms of a single, continuous variable. This caricature [...] falls short for placement and instruction problems based on students' internal representations of systems, problem-solving strategies, or reconfiguration of knowledge as they learn [...] Educational measurement faces today a crisis that would appear to threaten its very foundations [Mislevy in Frederiksen *et al.* 1993, S. 19].

3.5 Bayes-Inversion und Bevölkerungsmodell

Die Wahrscheinlichkeiten, dass bestimmte Probanden auf bestimmte Aufgaben bestimmte Antworten geben (1), kann man zusammensetzen zu der Wahrscheinlichkeit, dass der Test als ganzer ein bestimmtes Antwortmuster $\underline{\underline{\kappa}}$ liefert:

$$P(\underline{\underline{\kappa}}|\underline{\underline{\pi}}, \underline{\theta}) = \prod_{v\in\mathcal{V}} \prod_{i\in\mathcal{I}_v} A_i(\kappa_{vi}, \underline{\pi}_i, \theta_v)\,. \tag{8}$$

Dabei steht $\underline{\underline{\pi}}$ für die Gesamtheit aller Aufgabenparameter und $\underline{\theta}$ für die Kompetenzen sämtlicher Probanden. Die Skalierung stellt somit ein Umkehrproblem:

das Antwortmodell liefert $\underline{\underline{\kappa}}$ für gegebene $\underline{\underline{\pi}}$ und $\underline{\theta}$; tatsächlich ist aber $\underline{\underline{\kappa}}$ gegeben, und $\underline{\underline{\pi}}$ und $\underline{\theta}$ sind gesucht. Um dieses Problem zu lösen, wird die bedingte Wahrscheinlichkeit (8) mit Hilfe des Satzes von Bayes invertiert:

$$P(\underline{\underline{\pi}}, \underline{\theta}|\underline{\underline{\kappa}}) = \frac{P(\underline{\underline{\kappa}}|\underline{\underline{\pi}}, \underline{\theta})P(\underline{\underline{\pi}}, \underline{\theta})}{P(\underline{\underline{\kappa}})} \,. \tag{9}$$

Hier wird deutlich, dass eine Testauswertung mit der probabilistischen Testtheorie nicht bestimmte Werte für Aufgabenparameter und Kompetenzen, sondern Wahrscheinlichkeitsdichten liefert – in Form von bedingten Wahrscheinlichkeiten, die von den vorliegenden Testantworten $\underline{\underline{\kappa}}$ abhängen.

Die rechte Seite enthält neben dem Antwortmodell $P(\underline{\underline{\kappa}}|\underline{\underline{\pi}}, \underline{\theta})$ auch noch zwei nicht bedingte Wahrscheinlichkeiten. Der Nenner $P(\underline{\underline{\kappa}})$ geht aus den Modellannahmen hervor, indem man den Zähler über alle Komponenten von $\underline{\underline{\pi}}$ und $\underline{\theta}$ integriert. Der Faktor $P(\underline{\underline{\pi}}, \underline{\theta})$ im Zähler ermöglicht es, a-priori-Annahmen über die Wahrscheinlichkeit bestimmter Aufgabenparameter oder Kompetenzwerte zu berücksichtigen. Über die Aufgabenparameter werden in PISA keine a-priori-Annahmen getroffen, es wird also $P(\underline{\underline{\pi}}, \underline{\theta}) = P(\underline{\theta})$ gesetzt. Für die Kompetenzwerte wird eine Normalverteilung $\mathcal{N}$ mit Mittelwert 0 und Standardabweichung δ angenommen. Im weiteren Verlauf der Auswertung wird $P(\underline{\theta})$ als eine bedingte Wahrscheinlichkeit

$$P(\underline{\theta}|\delta) = \prod_{v \in \mathcal{V}} \mathcal{N}(\theta_v;\, 0,\, \delta) \tag{10}$$

aufgefasst und δ als freier Parameter so angepasst, dass die Gesamtheit von Antwort- und Bevölkerungsmodell eine möglichst wahrscheinliche Beschreibung der empirischen Daten $\underline{\underline{\kappa}}$ liefert.

Für eine solche parametrische a-priori-Verteilung (*latent distribution*) gibt es keine theoretische Rechtfertigung. Es gibt auch kaum Literatur dazu. Die allermeisten Autoren, die die Testtheorie vorangetrieben haben, sind an psychologischen, nicht an soziologischen Fragen interessiert (Bock 1997); in typischen Anwendungen geht es um Diagnose und Prognose für Individuen, nicht um die vergleichende Untersuchung von Merkmalsverteilungen in Subpopulationen. Bildungstestkonstrukteure wie Adams sind in der Psychometrie allenfalls Außenseiter; ihre Arbeiten werden dort kaum rezipiert. Eine kritische Auseinandersetzung mit der Anwendung der probabilistischen Testtheorie in Bildungsstudien hat bisher kaum stattgefunden: die Psychometriker haben kein Interesse, die Pädagogen keinen Zugang.

In dem Buch über Rasch-Modelle von Fischer und Molenaar (1995) werden latente Verteilungen nur ganz am Rande erwähnt (S. 47f., 254, 280f.); eine Begründung ist diesen Passagen nicht zu entnehmen. In der Literatur (so auch in Adams *et al.* 1997b) wird regelmäßig auf Bock und Aitkin (1981) zurückverwiesen, aber auch dort wird die simultane Schätzung von Schwierigkeitsparametern und Bevölkerungsverteilungsparameter nicht theoretisch gerechtfertigt, sondern

nur auf numerische Beispiele gestützt, die suggerieren, man könne sich auf diese Weise von einer groben Schätzung der Bevölkerungsverteilung zu einem „“empirical” prior“ vorarbeiten. Die Anführungszeichen um „empirical“ stehen original bei Bock und Aitkin – denn die Idee, eine Bayes'sche a-priori-Verteilung empirisch nachzubessern, wirkt in der Tat inkonsistent.

Rasch selbst war strikt dagegen, sein Antwortmodell um die Annahme zu erweitern, latente Fähigkeiten seien in irgendeiner Population normalverteilt. Sein Schüler Andersen berichtet:

> Georg Rasch had a very obvious animosity towards the normal distribution. At certain occasions [...] he would invite all persons present to a party on his front lawn to burn all books containing the word “normal distribution”. This animosity came from two applications of the normal distribution, which Georg Rasch felt was completely unjustified. The first one was standardization of educational or psychological tests, in particular intelligence tests, where individuals were classified by their position in terms of percentiles of a normal distribution describing the variation of the test score over a “standard population” [...] He was aware that it was important to be able to compare populations, or at least groups of individuals, for example, at different points in time. But it never took the form of a standardization [in Fischer/Molenaar 1995, S. 385].

Bock hat sich demgegenüber jahrzehntelang für die Vorgabe einer Normalverteilung oder einer anderen Fähigkeitsverteilung in der Parameterschätzung eingesetzt, meint aber, dass eine Normierung auf eine Standardpopulation zwar für Einstufungstests, weniger jedoch für das Qualitätsmonitoring im Bildungswesen vernünftig sei (1997, S. 30). Glas (2005) weist darauf hin, dass die Methodologie zur Überprüfung der Gültigkeit der Modellannahmen unterentwickelt ist.

Es gibt keinen Grund für die Annahme, dass Schülerkompetenzen normalverteilt sind. Bei allen möglichen psychometrischen Merkmalen sind Abweichungen von der Normalverteilung der Normalfall (Walberg *et al.* 1984; Micceri 1989). Solche Abweichungen aber können Maximum-Likelihood-Schätzungen mit normalverteiltem Prior stark verzerren (Molenaar in Fischer/Molenaar 1995, S. 48). Sie führen insbesondere dazu, dass die Schätzung extremer Parameterwerte (circa zwei Standardabweichungen außerhalb vom Zentrum) „in nichttrivialer Weise“ verzerrt wird (Woods/Thissen 2006). Demnach sind Aussagen über „Risikogruppen“ mit extrem schlechten PISA-Ergebnissen auch aus messtheoretischer Sicht völlig ungesichert.

3.6 Schätzung der Aufgabenparameter

Die Wahrscheinlichkeitsdichte einzelner Aufgabenparameter oder einzelner Kompetenzwerte erhält man aus (9), indem man alle übrigen Parameter ausintegriert („marginalisiert“). Weil im einen Fall über sehr viele Schüler, im anderen über

recht wenige Aufgaben gemittelt wird, haben die Wahrscheinlichkeitsdichten der Aufgabenschwierigkeiten erheblich geringere Varianzen als die der Schülerkompetenzen. Daher ist es vermutlich für viele Zwecke ausreichend und angemessen, den probabilistischen Charakter der Aufgabenparameter zu vernachlässigen und nur Kompetenzen durch Wahrscheinlichkeitsdichten zu beschreiben.

In PISA werden, wie oben skizziert, in dem „internationale Kalibrierung“ genannten Auswerteschritt die wahrscheinlichsten Werte der Aufgabenschwierigkeiten zusammen mit dem Parameter δ des Bevölkerungsmodells bestimmt.[26] Da für $\underline{\underline{\pi}}$ und δ keine a-priori-Verteilung gegeben ist, liefert der Satz von Bayes einfach

$$P(\underline{\underline{\pi}}, \delta|\underline{\underline{\kappa}}) = \frac{P(\underline{\underline{\kappa}}|\underline{\underline{\pi}}, \delta)}{P(\underline{\underline{\kappa}})} \,. \tag{11}$$

Zur Maximierung dieser Wahrscheinlichkeit setzt man am bequemsten die Ableitungen des Logarithmus von

$$P(\underline{\underline{\kappa}}|\underline{\underline{\pi}}, \delta) = \prod_{v\in\mathcal{V}} \int \mathrm{d}\theta_v \, P(\underline{\underline{\kappa}}|\underline{\underline{\pi}}, \underline{\theta}) P(\underline{\theta}|\delta) \tag{12}$$

$$= \prod_{v\in\mathcal{V}} \int \mathrm{d}\theta \prod_{i\in\mathcal{I}_v} A_i(\kappa_{vi}, \underline{\pi}_i, \theta) \mathcal{N}(\theta;\, 0,\, \delta) \tag{13}$$

nach δ und nach den Komponenten von $\underline{\underline{\pi}}$ gleich Null. Man erhält ein nichtlineares Gleichungssystem, das numerisch gelöst werden muss. Die dazu üblicherweise eingesetzte Iteration (Bock/Aitkin 1981, detailliert beschrieben und begründet von Woodruff/Hanson 1997) ist ein Spezialfall des Estimation-Maximization-Algorithmus, der recht robuste Konvergenzeigenschaften hat, aber sehr langsam sein kann (McLachlan/Krishnan 1997).

ACER berücksichtigt in der „internationalen Skalierung“, also der Bestimmung von $\underline{\underline{\pi}}$ und δ, nicht den vollen Rohdatensatz, sondern nur 500 Probanden pro OECD-Staat (TR, S. 128). Ein Grund wird nicht genannt. Eine denkbare Erklärung sind Rechenzeitprobleme infolge der langsamen Konvergenz: solange man nicht bestimmte speicherorganisatorisch aufwändige Optimierungen vornimmt, ist der Rechenaufwand zur Bestimmung von $\underline{\underline{\pi}}$ und δ im wesentlichen proportional zur Anzahl der berücksichtigten Probanden. Diesem Erklärungsversuch steht allerdings Köllers Mitteilung entgegen, ConQuest gelte aktuell als eine „äußerst leistungsstarke Software“ (Anh. D).

[26] An dieser Stelle bricht die Beschreibung des Skalierungsverfahrens im Technischen Bericht plötzlich ab. Nachdem bis hierhin lange, unnötig allgemein gehaltene Passagen weitgehend wörtlich aus dem ConQuest-Handbuch übernommen wurden, wird der weitere Gang der internationalen Kalibrierung in einem einzigen Satz zusammengefasst (TR, S. 122). Genau an dieser Bruchstelle hatte ich in W1 den Technischen Bericht missverstanden und die Skalierung falsch rekonstruiert.

3.7 Konditionierung mit Hintergrundvariablen

Nach Festlegung der Aufgabenparameter $\underline{\underline{\pi}}$ und des Bevölkerungsverteilungs-Parameters δ erhält man die Kompetenzen jedes einzelnen Probanden v – allerdings nicht als scharfen Zahlenwert, sondern als Wahrscheinlichkeitsverteilung

$$P_v(\theta_v) \equiv P(\theta_v|\underline{\underline{\pi}}, \underline{\kappa}_v, \delta) \tag{14}$$

$$= \frac{P(\underline{\kappa}_v|\underline{\underline{\pi}}, \theta_v)P(\theta_v|\delta)}{P(\underline{\kappa}_v|\underline{\underline{\pi}}, \delta)} \tag{15}$$

$$= \frac{\prod_{i\in\mathbb{J}_v} A_i(\kappa_{vi}, \underline{\pi}_i, \theta_v)\mathcal{N}(\theta_v;\, 0,\, \delta)}{\int \mathrm{d}\theta' \prod_{i\in\mathbb{J}_v} A_i(\kappa_{vi}, \underline{\pi}_i, \theta')\mathcal{N}(\theta';\, 0,\, \delta)}\,. \tag{16}$$

Diese Verteilung ist umso schmaler, je modellkonformer ein Proband den Test bearbeitet hat; aber auch bei perfekt modellgemäßem Verhalten, wie man es auf dem Computer simulieren kann, hat P_v aufgrund der probabilistischen Natur der zugrundeliegenden Modelle und der begrenzten Aufgabenanzahl noch eine endliche Breite.

Um die Breite der P_v und damit die Standardfehler später zu berechnender Populationsstatistiken zu verringern, wird das Bevölkerungsmodell gegenüber der hier angegebenen, für alle Probanden gleichen Normalverteilung durch Einbeziehung konditionierender Hintergrundvariablen verfeinert. Eine dieser Variablen ist der schulweite Mathematikkompetenzmittelwert. Dreizehn Variablen, jeweils mit dem Wertebereich $\{0, 1\}$, geben an, welches Testheft der Proband bearbeitet hat (TR, S. 408). Alle übrigen Variablen werden dem *Student Questionnaire* entnommen. Davon werden aber nur das Geschlecht des Probanden und die Berufe seiner Eltern in international einheitlicher Weise berücksichtigt. Die übrigen Schülerantworten werden einer Hauptkomponentenanalyse unterworfen, um Linearkombinationen auszuwählen, die zusammengenommen 95 % der Varianz der Originalvariablen erklären. Diese Analyse wird im Technischen Bericht nur kursorisch beschrieben (S. 129); es wird nicht einmal auf Anhang 10 verwiesen, der die Binärkodierung der Questionnaire-Daten angibt. Offensichtlich hat ACER wenig Vertrauen in die internationale Vergleichbarkeit der Schülerantworten, denn die Hauptkomponentenanalyse wird nach Staaten getrennt durchgeführt. Eigenwerte und Eigenvektoren sind nicht dokumentiert.

Sei $\mathcal{M}_S$ die Gesamtmenge der für einen Staat S ausgewählten Merkmale (international einheitliche Hintergrundvariablen und Linearkombinationen aus der Hauptkomponentenanalyse). Sei h_{vm} der Zahlenwert eines solchen Merkmals $m \in \mathcal{M}_S$ für einen Probanden v. Für die Bestimmung der Schülerkompetenzen wird das Antwortmodell erneut an die kognitiven Testergebnisse $\underline{\underline{\kappa}}$ angepasst, diesmal aber separat für jeden Staat, mit festgehaltenen Aufgabenschwierigkeiten $\underline{\underline{\pi}}$ und mit einem Bevölkerungsmodell, das gegenüber (10) um eine Vielzahl

von Parametern erweitert ist:

$$\mathcal{N}\left(\theta_v;\ \sum_{m\in\mathcal{M}_S}\eta_{Sm}h_{vm},\ \delta_S\right). \tag{17}$$

Die Koeffizienten η_{Sm} geben an, wie stark die einzelnen Hintergrundmerkmale h_{vm} den Mittelwert der Normalverteilung $\mathcal{N}$ verschieben. Die wahrscheinlichsten Werte für diese Koeffizienten (zusammengefasst: $\underline{\eta}_S$) und für die Breite δ_S der Verteilung werden mit der Methode aus 3.6 bestimmt. Die Zahlenwerte der $\underline{\eta}_S$ und δ_S sind nicht dokumentiert. Das so präzisierte Bevölkerungsmodell wird in Gleichung (16) eingesetzt und der weiteren Auswertung zu Grunde gelegt.[27]

Warum ein so kompliziertes Vorgehen gewählt wurde, wird im Technischen Bericht nicht begründet – es wird nur erwähnt, dass es in TIMSS und der amerikanischen NAEP-Studie auch schon so gemacht wurde, und auf Adams *et al.* (1997a) verwiesen. Dort werden vier mögliche Vorteile einer differenzierten Bevölkerungsmodellierung genannt. Drei davon sind für PISA nicht einschlägig. Es bleibt als einziger erkennbarer Vorteil eine kleinere Breite der Wahrscheinlichkeitsdichten $P_v(\theta_v)$ und damit ein geringerer Standardfehler von Populationsstatistiken – wodurch die Schwelle sinkt, ab wann sich ein Unterschied zwischen zwei Populationen von der Nullhypothese abhebt.

Damit wird klar, warum bei der Wahl der Merkmalskombinationen $\mathcal{M}_S$ nach Staaten und nicht nach anderen denkbaren Einteilungen der OECD-Stichprobe unterschieden wird: die konditionierenden Hintergrundvariablen dienen primär dem Zweck, geringe Unterschiede zwischen Staaten als signifikant deklarieren zu können. Der Preis dafür ist eine unverhältnismäßige Komplexität unzureichend dokumentierter Prozeduren.

3.8 Statistische Auswertungen und offizielle Standardfehler

Ziel aller Auswertungen sind statistische Aussagen über die Häufigkeitsverteilung der Kompetenzen in bestimmten Substichproben $\mathcal{S}\subset\mathcal{V}$, die, bei Berücksichtigung der Probandengewichte w_v, für Subpopulationen der OECD stehen.[28] Wären die Probandenkompetenzen θ_v exakt bekannt, würde man konventionelle Statistiken

$$\overline{T}=\sum_{v\in\mathcal{S}}w_vT(\theta_v,h_v) \tag{18}$$

[27] Möglicherweise wird diese Analyse auch simultan für alle vier oder sieben Testgebiete $\mathcal{G}$ durchgeführt; dann hängen θ und η auch noch von $g\in\mathcal{G}$ ab, und laut Technischem Bericht ist δ^2 durch eine Kovarianzmatrix zu ersetzen.

[28] In W1 (S. 145) hatte ich zu Unrecht kritisiert, dass bei der Skalierung alle Probanden gleich gewichtet werden. Es ist korrekt, die w_v erst in (18) zu berücksichtigen.

berechnen, wobei die Normierung $\sum_{v\in\mathcal{S}} w_v = 1$ vorausgesetzt werden soll. Die gebräuchlichsten Statistiken sind Mittelwert $[T_E(\theta_v) = \theta_v]$ und Varianz $[T_V(\theta_v) = (\theta_v - \overline{T_E}\,)^2]$; in weiterführenden Auswertungen interessiert man sich zum Beispiel für die Korrelation zwischen dem kognitiven Testergebnis θ_v und einer Hintergrundvariablen h_v.

Da aber die θ_v nicht exakt bekannt sind, können auch Statistiken nur in Form von Wahrscheinlichkeitsaussagen ausgewertet werden, basierend auf den Wahrscheinlichkeitsverteilungen $P_v(\theta_v)$ aus (16). So erhält man anstelle von $\overline{T}$ nur einen Erwartungswert dieses Mittelwerts,

$$\langle\overline{T}\rangle \equiv \prod_{v\in\mathcal{S}} \int \mathrm{d}\theta_v\, P_v(\theta_v)\, \overline{T}\,, \tag{19}$$

der bequemer als Mittelwert von Erwartungswerten

$$\langle\overline{T}\rangle = \cdots = \sum_{v\in\mathcal{S}} w_v \int \mathrm{d}\theta_v\, P_v(\theta_v) T(\theta_v, h_v) \equiv \sum_{v\in\mathcal{S}} w_v \langle T_v\rangle \tag{20}$$

berechnet werden kann.

An dieser, und nur an dieser Stelle werden in der offiziellen Auswertung Messunsicherheiten abgeschätzt. Die in den Ergebnisberichten angegebenen *Standardfehler*, im folgenden als $u_{\overline{T},\mathrm{PISA}}$ bezeichnet,[29] werden durch Addition von zwei Varianzen berechnet (TR, S. 131, Formel 9.18; DAM, S. 79):[30]

$$u_{\overline{T},\mathrm{PISA}} = \sqrt{{u_{\overline{T},P}}^2 + {u_{\overline{T},w}}^2}\,. \tag{21}$$

Der Beitrag $u_{\overline{T},P}$ ist durch die endliche Breite der Wahrscheinlichkeitsverteilung $\prod_v P_v(\theta_v)$ verursacht. Mit etwas Rechnung findet man

$${u_{\overline{T},P}}^2 = \left\langle \left(\overline{T} - \langle\overline{T}\rangle\right)^2 \right\rangle = \cdots = \sum_{v\in\mathcal{S}} {w_v}^2 \left\langle \left(T_v - \langle T_v\rangle\right)^2 \right\rangle\,. \tag{22}$$

Da die normierten Gewichte w_v im Mittel invers proportional zur Stichprobengröße $|\mathcal{S}|$ sind, bewirkt ihr quadratisches Auftreten in (22), dass $u_{\overline{T},P}$ asymptotisch proportional zu $|\mathcal{S}|^{-1/2}$ ist; mit zunehmender Stichprobengröße nimmt diese Unsicherheit also langsam ab.

Der zweite Beitrag zu den offiziellen Standardfehlern ist die *sampling variance* ${u_{\overline{T},w}}^2$, die die Unsicherheit der Stichprobenziehung und damit der Probandengewichte w_v angibt (TR, Kapitel 8). Aufgrund der Komplexität der mehrstufigen Stichprobenziehung ist eine formelmäßige Herleitung der $u_{\overline{T},w}$ angeblich

[29] Um das überstrapazierte Formelzeichen σ zu vermeiden, das im Technischen Bericht auch für die hier δ genannte Breite des Bevölkerungsmodells steht.

[30] Die Berücksichtigung nur dieser beiden Varianzen wurde schon beim amerikanischen NAEP als unzureichend krititisiert (Yamamoto/Mazzeo 1992).

nicht möglich (DAM, S. 44). Stattdessen wird die diesbezügliche Varianz von (20) in einem Monte-Carlo-Verfahren abgeschätzt, wofür im internationalen Datensatz pro Proband achtzig modizifizierte Repliken $w_v^{(k)}$ abgelegt sind, die die Unsicherheit der w_v widerspiegeln sollen.

Der deutsche Bericht dokumentiert die $u_{\overline{T},\mathrm{PISA}}$ besonders gründlich, nämlich nicht nur für die Mittelwerte $\overline{T_E}$, sondern auch für die Standardabweichungen $\overline{T_V}^{1/2}$ der nationalen Kompetenzverteilungen (D03b, S. 70, 99, 118, 157). In den offiziellen Auswertungen sind die $u_{\overline{T},\mathrm{PISA}}$ unentbehrlich für die Interpretation aller Statistiken $\overline{T}$, denn sie allein entscheiden darüber, ab wann ein Unterschied zwischen zwei Subpopulationen als *signifikant* angesehen wird. Von anderen Fehlerquellen, die nicht modellimmanent quantifiziert werden können, ist nirgendwo ernsthaft die Rede.

Um qualitativ zu verstehen, wodurch in dieser offiziellen Sichtweise die Messgenauigkeit von PISA begrenzt wird, wäre es nötig, zu erfahren, in welchem Zahlenverhältnis die beiden Beiträge zu $u_{\overline{T},\mathrm{PISA}}$ stehen. Auch diese Angabe ist nicht leicht zu finden. Ein Zahlenbeispiel (DAM, S. 97 f.) deutet darauf hin, dass ${u_{\overline{T},P}}^2$ gegenüber ${u_{\overline{T},w}}^2$ völlig vernachlässigbar ist. Das bedeutet: für die Produktion „signifikanter“ Rangunterschiede ist es unerheblich, ob die einzelnen Testaufgaben mehr oder weniger modellgerecht funktionieren; die statistische Genauigkeit der Kompetenzwerte ist allein durch die schwierige, mehrstufige Stichprobenziehung begrenzt. Nebenbei erklärt das, warum die Standardfehler in Island und Luxemburg, wo der ganze Altersjahrgang getestet wurde, besonders niedrig sind.

3.9 „Plausible“ Kompetenzwerte

Im vorigen Abschnitt wurde offengelassen, wie die Integrale in (19) oder (20) berechnet werden. Im Prinzip ist das ein völlig unkritisches technisches Detail; man muss nur für jeden Probanden v an hinreichend vielen Stützpunkte θ_v die Wahrscheinlichkeitsdichte $P_v(\theta_v)$ berechnen, dann kann man die Integrale mit beliebiger Genauigkeit durch Summen approximieren. Problematisch ist das nur für die Kommunikation: Die Untersuchung von Statistiken $T(\theta_v, h_v)$ soll auch unabhängig vom Konsortium arbeitenden Sekundärauswertern ermöglicht werden, die sich zum Beispiel für die Korrelation zwischen θ_v und einer neu konstruierten Hintergrundvariablen h_v interessieren. Dafür müssten Hunderttausende Wahrscheinlichkeitsdichten $P_v(\theta_v)$ mit hinlänglicher Genauigkeit verfügbar gemacht werden. Dieser Aufwand wird in PISA durch ein Monte-Carlo-Verfahren vermieden: es genügen wenige Stützpunkte pro Proband, wenn diese *zufällig* gezogen werden. Diese Stützpunkte werden als *plausible Werte* bezeichnet und als Teil des internationalen Datensatzes veröffentlicht.

Allerdings hat man sich damit ein Kommunikationsproblem auf einer anderen Ebene eingehandelt: manche Projektbeteiligte verstehen die Methode nicht

und schreiben ihr Wirkungen zu, die sie, als eine rein numerische Approximationstechnik, nicht haben kann. Olaf Köller verteidigt die Verwendung eines Ein-Parameter-Modells mit den Vorzügen plausibler Werte (Anhang D), und die deutsche Pisa-Expertengruppe Mathematik (Anhang E) meint, durch plausible Werte könnten Schätzfehler verringert werden.

Um solchen Missverständnisse entgegenzutreten, sei wiederholt: die Methode der plausiblen Werte ist nicht mehr als ein Monte-Carlo-Verfahren zur Berechnung der hochdimensionalen Integrale (19) bzw. (20). Bei großem Stichprobenumfang $|\mathcal{S}|$ genügt es, pro Integral eine recht geringe Anzahl J von Stützpunkten $\theta_v^{(j)}$ (mit $j = 1, \ldots, J$) zu berücksichtigen,

$$\sum_{v\in\mathcal{S}} w_v \int \mathrm{d}\theta_v \, P_v(\theta_v) T(\theta_v, h_v) \simeq \sum_{v\in\mathcal{S}} w_v \, \frac{1}{J} \sum_{j\le J} T(\theta_v^{(j)}, h_v) \,, \tag{23}$$

sofern die Stützpunkte ohne Zurücklegen unter jedesmaliger Berücksichtigung der Wahrscheinlichkeitsdichte P_v gezogen werden.[31] In PISA wurde $J = 5$ für ausreichend befunden: ACER zieht fünf Stützpunkte $\theta_v^{(j)}$ pro Proband und Testgebiet und fügt diese als „plausible values“ dem internationalen Datensatz bei. Sekundärauswerter haben keinen Zugriff auf die P_v, können aber in guter Näherung Integrale der Form (20) berechnen, indem sie die plausiblen Werte $\theta_v^{(j)}$ in die rechte Seite von (23) einsetzen.

3.10 Synthetische Kompetenzwerte

In PISA 2000 wurden den Probanden nur in denjenigen Testgebieten Kompetenzwerte zugeordnet, in denen sie Aufgaben bearbeitet hatten. Deshalb mussten den Probanden je nach Testgebiet auch unterschiedliche statistische Gewichte zugeschrieben werden. Zur Vereinfachung der Datenstruktur und zur Verringerung stochastischer Standardfehler haben Wu *et al.* in einem Konferenzbeitrag (2002) vorgeschlagen, die hohe Korrelation zwischen den verschiedenen Testgebieten auszunutzen, um fehlende Kompetenzwerte zu *schätzen*. Sie antizipierten zwar Vermittlungsprobleme:

> On the other hand, we will need to avoid possible violations of policy and ethical standards, as well as convince the less scientific community, that we can obtain reasonable estimates of students' scores given partial information [...] If the method of imputation is adopted, we do need, however, to ensure that such methods are acceptable politically, as we would appear to the less scientific community that we are producing score estimates when students did not even attempt any items [S. 25 f.];

[31] Im fiktiven Grenzfall $J \to \infty$ würde die Häufigkeitsverteilung der $\theta_v^{(j)}$ mit der vorgegebenen Verteilung P_v übereinstimmen.

nichtsdestoweniger wurde dieser Vorschlag ohne weitere wissenschaftliche Debatte in PISA 2003 umgesetzt.[32]

In PISA 2003 waren beispielsweise Naturwissenschaftsaufgaben in nur sieben von dreizehn regulären Testheften enthalten (Tab. 2). Nichtsdestoweniger werden auch für die Probanden, die eines der übrigen sechs Hefte bearbeitet haben, mit dem oben beschriebenen Verfahren Naturwissenschaftskompetenzen geschätzt. Gleichung (16) vereinfacht sich dabei radikal: weil $\mathfrak{I}_v = \emptyset$, ist die Wahrscheinlichkeitsdichte $P_v(\theta_v)$ allein durch das Bevölkerungsmodell gegeben, in das gemäß der Modifikation (17) je nach Staat S bestimmte Hintergrundmerkmale $\mathcal{M}_S$ eingerechnet sind. Kognitive Testleistungen werden dabei nur halbherzig, nämlich über die auf Schulebene gemittelte Mathematikkompetenz, berücksichtigt. Die unterschiedliche Schwierigkeit der Testhefte wird auf nationaler Ebene berücksichtigt; dabei wird Heft 9 als das einzige, das Aufgaben aus allen vier Testgebieten enthält, zum Bezugspunkt gemacht.

Einer vermutlich falsch beschrifteten Tabellenspalte ist zu entnehmen, dass der Standardfehler der nationalen Kompetenzmittelwerte typischerweise um 10 % geringer ist, als er wäre, wenn nur die 7/13 aller Probanden berücksichtigt worden wären, die tatsächlich in Naturwissenschaften getestet wurden.[33]

> It is important to realise that this is not an artificial result that is merely due to an increase in sample size, but is a genuine reduction in the error caused by the increase in the total available information about the proficiency distribution [TR, S .207].

Um mich von der „less scientific community“ abzugrenzen, behaupte ich zu verstehen, dass diese creatio ex nihilo bei hinreichend idealem Datenmaterial mindestens korrekte Mittelwerte liefert. Allerdings bleibt mir unklar, wie weit der Gültigkeitsbereich der fiktiven Kompetenzwerte reicht und wie garantiert werden kann, dass es in Sekundärauswertungen nicht doch zu Artefakten kommt.

Für die realen Daten aus PISA ist die Verankerung der Schätzungen an einem bestimmten Testheft jedenfalls nicht zu rechtfertigen. Die Lösungshäufigkeiten der einzelnen Aufgabenblöcke unterscheiden sich sowohl von Staat zu Staat als auch von Heft zu Heft ganz erheblich. Beispielsweise fallen in Griechenland die Mathematikaufgaben aus Heft 9 besonders schwer: die Schüler, die dieses Heft zu bearbeiten hatten, haben im Mittel nur 423 Kompetenzpunkte, gegenüber 445 Punkten im nationalen Durchschnitt. Andererseits laufen die

[32] Der Technische Bericht (TR, S. 129 und identisch S. 206) verweist nicht auf begutachtete Fachliteratur, sondern nur auf die technischen Berichte zur amerikanischen NAEP-Studie und zu TIMSS. Zumindest in letztgenannter Quelle (Macaskill *et al.* in Martin/Kelly 1998) wird die in PISA 2003 vorgenommene Ausweitung der Kompetenzwertschätzung *nicht* begründet.

[33] In der Tabelle (TR, S. 209, Tab. 13.23, letzte Spalte; Erläuterung S. 207) ist die Änderung der *Varianz* angegeben, die typischerweise 20 % beträgt.

Naturwissenschaftsaufgaben aus Heft 9 vergleichsweise gut: 494 Punkte gegenüber 465 im Mittel über die sieben einschlägigen Hefte. Infolge dieser schiefen Verankerung wird denjenigen Schülern, die keine Naturwissenschaftsaufgaben zu bearbeiten hatten, eine mittlere Naturwissenschaftskompetenz von 499 zugeschrieben (Neuwirth in Neuwirth *et al.* 2006, S. 53). Im Mittel über alle dreizehn Hefte ergibt sich so für Griechenland der nach außen mitgeteilte Kompetenzwert 481 (LTW, S. 294), der um 16 *(sechzehn!)* Punkte über dem empirischen Wert 465 liegt. Auch die offiziellen Ergebnisse für Mexiko und Japan sind stark noch oben verschoben (um 12 bzw. 11 Punkte). Die kanadischen und dänischen Naturwissenschaftsleistungen werden hingegen um 9 bzw. 11 Punkte zu niedrig angegeben; für Kanada wird das sogar in Klartext mitgeteilt (TR, S. 211). Für die Hälfte aller OECD-Staaten beträgt die Verzerrung 4 oder mehr Punkte. Im Lesetest ist der „Testheft-9-Effekt“ generell etwas schwächer; die Verzerrungen liegen zwischen +12 (Griechenland) und −9 (Dänemark, Japan).

Diese Befunde sind so unglaublich, dass ich ohne die Bestätigung durch Neuwirth (*loc. cit.*) nicht für möglich gehalten hätte, dass sie nicht bloß einen weiteren Dokumentationsfehler darstellen: um *stochastische* Standardfehler, die durchweg weniger als 5 Punkte betragen (LTW, S. 294), um typischerweise 10 % zu verringern, nimmt ACER *systematische* Verfälschungen der empirischen Daten um bis zu 16 Punkte in Kauf.

3.11 Nachträgliche Umskalierung

In den internationalen und nationalen Ergebnisberichten werden Aufgabenschwierigkeiten und Schülerkompetenzen auf einer Punkteskala mitgeteilt, die so konstruiert ist, dass die Kompetenzwerte OECD-weit den Mittelwert 500 und die Standardabweichung 100 haben („PISA scale“). Intern findet die Auswertung jedoch auf einer anderen Skala statt, die durch die oben beschriebene Modellierung festgelegt ist. Auf dieser Skala („logits“) liegt das Zentrum der Kompetenzverteilung ungefähr bei 0, und die Standardabweichung ist von der Größenordnung 1. Der Technische Bericht springt unsystematisch und ohne Erklärung zwischen beiden Skalen hin und her, besonders auffällig in Kapitel 13, wo es mit Tab. 13.19 f. sogar eine Bilingue gibt.

Diese Umskalierung ist im Technischen Bericht mangelhaft dokumentiert; sie wird erst am Ende des Ergebniskapitels 13 angegeben, und die Herkunft der Koeffizienten wird nicht erläutert.[34] Für den Mathematiktest lautet die Umskalierung

$$P = 500 + 100(J + 0{,}1344)/1{,}2838 = 510{,}47 + 77{,}89J\,. \tag{24}$$

[34]Das hat erheblich zur unzutreffenden Rekonstruktion der Skalierung in W1 beigetragen; siehe dazu Anhang B.

Die Formelbuchstaben werden nicht eingeführt; ich deute J als interne Kompetenzwerte θ und P als externe („PISA“) Werte, die ich im folgenden als θ^{P} notiere. Woher die Koeffizienten stammen, wird nicht erklärt; nur indirekt lässt sich erschließen, dass $-0{,}1344$ der Mittelwert und 1,2838 die Standardabweichung der OECD-weiten Häufigkeitsverteilung der internen Kompetenzwerte unter Einschluss der britischen Daten, unter Einschluss der Kurzhefte und unter Berücksichtigung der Probandengewichte sein muss.

Diese Umrechnung ist aber nicht die volle Wahrheit, denn sie kann nicht sowohl für Kompetenzwerte als auch für Aufgabenschwierigkeiten gelten. Abweichend von der im Technischen Bericht mitgeteilten Form es psychologischen Modelle (3) ff. stimmen Schwierigkeits- und Kompetenzwerte auf der nach außen mitgeteilten Skala dann überein, wenn die Wahrscheinlichkeit einer richtigen Lösung nicht 50 %, sondern 62 % beträgt.[35]

Das heißt: die Skalen, auf denen Kompetenz- und Schwierigkeitswerte nach außen kommuniziert werden, sind willkürlich gegeneinander verstimmt. Aufgabenschwierigkeiten werden *nicht* gemäß den Formeln aus Kapitel 13 des Technischen Berichts umgerechnet, sondern unterliegen einer zusätzlichen Verschiebung um $\ln(62/38)$. Beispielsweise sind die Schwierigkeiten von Mathematikaufgaben gemäß

$$\xi^{\mathrm{P}} = 510{,}47 + 77{,}89\big(\xi + \ln(62/38)\big) \tag{25}$$

umzurechnen. So nimmt das Rasch-Modell die in keinem offiziellen Bericht angegebene Gestalt

$$A^{\mathrm{P}}_{\mathrm{Rasch}}(\text{richtig}, \xi^{\mathrm{P}}, \theta^{\mathrm{P}}) = \frac{1}{1 + \exp\left[\frac{\xi^{\mathrm{P}} - \theta^{\mathrm{P}}}{77{,}89} - \ln\frac{62}{38}\right]} \tag{26}$$

an.

3.12 Kompetenzwerte sind im Grunde Punktsummen

Wenn man die konkret in PISA gewählten Antwortmodelle (3) ff. in (16) einsetzt und alle Faktoren abspaltet, die nicht von der Probandenkompetenz θ_v abhängen, nimmt deren Wahrscheinlichkeitsdichte eine überraschend einfache Form an. Ohne konditionierende Hintergrundvariable und ohne Normierung als Proportionalitätsbeziehung notiert:

$$P(\theta_v|\underline{\underline{\pi}}, \underline{\kappa}_v, \delta) \propto \prod_{i \in \mathcal{I}_v} A_i(\text{falsch}, \underline{\pi}_i, \theta_v) \exp(n_v \theta_v)\, \mathcal{N}(\theta_v;\, 0,\, \delta) \tag{27}$$

[35] LTW, S. 106, Endnote 5 zu S. 45. Dort wird behauptet, die Zahl 62 sei nicht willkürlich, sondern mit der Definition der Kompetenzstufen verknüpft. Allerdings beruht die Definition der Kompetenzstufen auf einem ganzen Bündel willkürlicher Festlegungen.

Die Details des Antwortmusters $\underline{\kappa}_v$ erweisen sich hier als irrelevant; sie gehen nur noch über eine Punktsumme n_v, die im wesentlichen die Anzahl richtig gelöster Aufgaben angibt, in die Kompetenzbewertung ein: n_v ist eine *suffiziente Statistik* für θ_v. Diese Eigenschaft gilt als entscheidender Vorteil der Rasch-Funktion gegenüber anderen Antwortmodellen (Molenaar in Fischer/Molenaar 1995, S. 10).

Solange man nur Probanden betrachtet, die dieselbe Aufgabenauswahl $\mathcal{I}_v$ zu bearbeiten hatten, stimmen auch die Vorfaktoren A_i und die in (27) nicht notierte Normierung überein. Dann wird allen Probanden, die dieselbe Punktsumme erzielt haben, exakt dieselbe Kompetenz-Wahrscheinlichkeitsdichte zugeschrieben. Die Umrechnung von Punktsummen in Kompetenzwerte ist zwar auch dann nicht trivial, doch man kann festhalten: Kompetenz wird in PISA allein über die Anzahl richtig gelöster Aufgaben gemessen.

Warum aber wird ein so aufwändiges und intransparentes Auswerteverfahren gewählt, wenn es letztlich auf eine einfache Punktsumme hinausläuft? Zwei Zwecke können eine Item-Response-Skalierung rechtfertigen:

(1) Probanden miteinander zu vergleichen, die *verschiedene* Aufgabensätze bearbeitet haben. Das kommt in PISA aus drei Gründen vor: (a) Verwendung verschiedener Testhefte; (b) Löschung einzelner Aufgaben in einzelnen Ländern, in einzelnen Heften oder für einzelne Probanden; (c) dreijährliche Wiederholung des Tests mit teils alten, teils neuen Aufgaben. Bei solchen Vergleichen unterscheiden sich die Vor- und Normierungsfaktoren in (27), so dass die gleiche Anzahl richtig gelöster Aufgaben bei Probanden, die verschiedene Hefte zu bearbeiten hatten, zu unterschiedlichen Kompetenzbeurteilungen führen kann. Aber das sind technische Details, die ein grundsätzliches Verständnis der Kompetenzbewertung zunächst eher behindern; sie ändern nichts daran, dass Kompetenzpunkte auf Ebene des einzelnen Probanden nichts als die Anzahl richtig gelöster Aufgaben wiederspiegeln. Insofern ist D. Lind (2004) unbedingt recht zu geben, dass man „die Rolle des Raschmodells bei PISA nicht überinterpretieren“ sollte.[36]

(2) Erst das „Zugeständnis“, dass Summenscore und Personenparameter hochgradig korreliert sind, bereitet laut Rost (2000) „den Weg für die Einsicht, worin denn der eigentliche Pfiff des Rasch-Modells liegt, nämlich in der Prüfung der Frage, ob man überhaupt einen Summenscore bilden darf“. Diese Prüfung wird ein konsistent negatives Ergebnis zeitigen (Teil 4).

Unabhängig von ihrer mathematischen Konsistenz mutet die Bewertung eines Tests durch bloßes Abzählen der Anzahl richtig gelöster Aufgaben, gemessen am Anspruch von PISA, reichlich primitiv an. Sie entspricht, jedenfalls in

[36] Cook *et al.* (1988) konstatieren, dass die Item-Response-Theorie und die klassische Testtheorie in ganz ähnlicher Weise scheitern, wenn die Grundannahme der eindimensionalen Bedingtheit des Schülerverhaltens verletzt ist – zum Beispiel durch uneinheitliche curriculare Validität der Aufgaben.

Deutschland, nicht der Erwartung von Schülern: Praxis in der Bewertung von Klassenarbeiten ist vielmehr, dass arbeitsaufwändige Aufgaben mehr Punkte bringen als schnell zu bearbeitende. Mangelnde Erfahrung mit PISA-ähnlichen Tests kann dazu führen, dass Schüler zu viel Zeit mit einzelnen, schwierigen oder langwierigen Aufgaben verbringen, statt die Gesamtzahl gelöster Aufgaben zu maximieren.

Überraschend ist auch, wie die Punktsumme n_v genau zustande kommt. Jede richtig gelöste dichotome Aufgabe trägt einen Punkt bei. Jede teilrichtig beantwortete trichotome Aufgabe trägt ebenfalls einen Punkt bei; eine voll richtige Antwort bringt bei diesen Aufgaben sogar zwei Punkte. Die richtige Lösung eines tetrachotomen Aufgaben trägt dementsprechend drei Punkte zu n_v bei. Dass einzelne Aufgaben mit doppeltem oder dreifachem Gewicht beitragen, nur weil dem Kodierer auch die Möglichkeit einer teilrichtigen Bewertung zur Verfügung stand, ist willkürlich und unplausibel. Auf diese Höhergewichtung einzelner Aufgaben wurden die Probanden auch nicht hingewiesen. Warum wurden die Partial-Credit-Modelle nicht so modifiziert, dass teilrichtige Lösungen mit halben oder drittel Punkten bewertet werden?

3.13 Die offizielle Skalierung ist nicht reproduzierbar

Die vorstehende Rekonstruktion der offiziellen Skalierung steht, wie eingangs gesagt, unter dem Vorbehalt, dass sie auf einer ungenauen und lückenhaften Dokumentation beruht. Sie kann auch nicht durch unabhängiges Nachprogrammieren und Vergleich der numerischen Ergebnisse verifiziert werden – und das nicht nur wegen der Komplexität der Rechnungen, wegen fehlender Vergleichsmöglichkeit mit unveröffentlichten Zwischenergebnissen (insbesondere den $\underline{\eta}_S$), wegen unklarer Details und wegen der inhärenten Verwendung von Zufallszahlen, sondern allein schon, weil elementarer *Input* nicht nachvollziehbar ist:

Spiegelbildlich zu (27) gilt im Rasch-Modell, dass die Anzahl der Probanden, die eine bestimmte Aufgabe i richtig gelöst haben, eine suffiziente Statistik für deren Schwierigkeitsbewertung ξ_i ist. Die Verteilung der Aufgaben auf mehrere Blöcke und die Verknüpfung von Antwort- und Bevölkerungsmodell verursachen Komplikationen, ändern aber nichts am Prinzip. Aus diesem Grund ist die Anzahl richtiger Lösungen oder, äquivalent dazu, deren relative Häufigkeit ρ_i eine essentielle Information nicht nur für didaktische Interpretationen auf der Ebene der einzelnen Aufgaben, sondern auch für Untersuchungen zur Skalierung.

Im Technischen Bericht sind die Lösungsquoten der einzelnen Aufgaben in den tabellarischen Anhängen 12–15 in einer mit „International % correct“ bezeichneten Spalte mitgeteilt. Unlogischerweise wird auch zu polychotomen Aufgaben nur *eine* Prozentzahl mitgeteilt. Wie die Prozentangaben zustande kommen, wird nicht erklärt. Die offizielle Auswertung legt zwei verschiedene

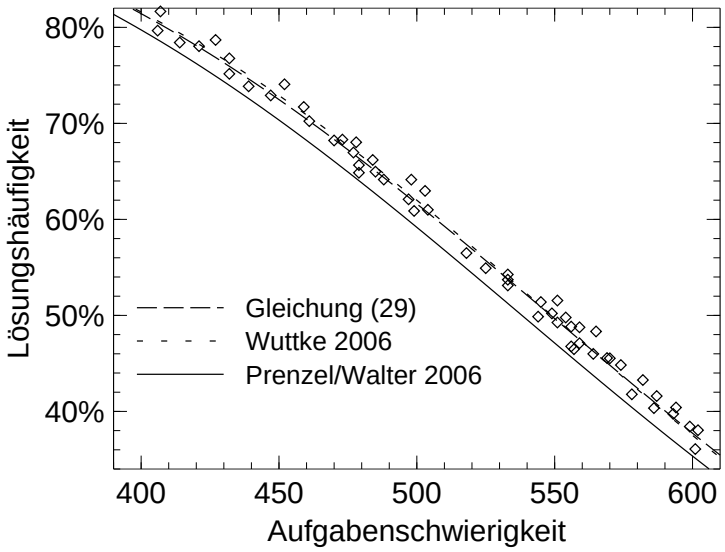

Abbildung 3: Lösungshäufigkeiten und Aufgabenschwierigkeiten der 76 dichotomen Mathematikaufgaben gemäß Technischem Bericht (TR, S. 412 f.). Die Streuung der Datenpunkte ist noch nicht erklärt. Nach derzeitigem Stand der Rekonstruktion scheint es, dass theoretisch der durch Gleichung (29) gegebene Verlauf zu erwarten wäre. Gezeigt ist außerdem die unzutreffende Rekonstruktion (30) aus W1 sowie die unzutreffende und empirisch deutlich unter den Datenpunkten liegende Rekonstruktion von Prenzel und Walter (2006).

Grundgesamtheiten nahe: (1) Wie bei der Kalibrierung der Aufgabenschwierigkeiten, also ohne nicht erreichte Aufgaben, ohne Sonderschüler, ohne Probandengewichte. Oder (2) umgekehrt, wie bei der Bestimmung der plausiblen Kompetenzwerte. Mit keiner dieser beiden Setzungen kann ich die Angaben aus dem Technischen Bericht reproduzieren: zum Beispiel ist die erste Mathematikaufgabe, „View Room“, laut Bericht zu 76,77 % gelöst worden; gemäß (1) finde ich 77,6 %, gemäß (2) 76,4 %. Bei anderen Aufgaben beträgt die Abweichung bis zu ±2 %.

In Abbildung 3 sind die ρ_i aus dem Technischen Bericht gegen die in denselben Anhängen 12–15 mitgeteilten Schwierigkeitsparameter ξ_i aufgetragen. Aus unerklärten Gründen ist der Zusammenhang nicht monoton: die ρ_i weichen um bis zu ±2 % von den Werten ab, die man bei gegebenen ξ_i aufgrund einer glatten Interpolation erwarten würde. Vielleicht beruhen diese Schwankungen auf der Zuordnung der Aufgaben zu unterschiedlich schwierigen Testheften, aber das ist wegen der Nichtreproduzierbarkeit der ρ_i nicht unabhängig nachprüfbar.

Mangelnde Reproduzierbarkeit wesentlicher Ergebnisse erschwert die kritische Auseinandersetzung mit einer Studie und stellt ihre Validität in Frage:

> Reproducibility is important in its own right, and is the standard for scientific discovery [Gentleman *et al.* 2004].

Das Problem, computergenerierte Ergebnisse in nachvollziehbarer Form zu dokumentieren, ist keineswegs auf PISA beschränkt, sondern eines der schwierigsten der heutigen Forschungskultur (Hothorn 2006). Eine Schere öffnet sich zwischen immer ausgefeilteren Methoden und einer oft nachlässigen Dokumentation, was zum Beispiel bei biomedizinischen Studien desaströse Folgen haben kann (Moher *et al.* 2004).

> Biased results from poorly designed and reported trials can mislead decision making in health care at all levels, from treatment decisions for the individual patient to formulation of national health policies. Critical appraisal of the quality of clinical trials is possible only if the design, conduct, and analysis [...] are thouroughly and accurately described in published articles [Altman *et al.* 2001].

Aber selbst wenn sie vollständig und fehlerfrei sein sollten, sind herkömmliche Berichte keine angemessene Publikationsform mehr:

> Indeed the problem occurs wherever traditional methods of scientific publication are used to describe computational research. In a traditional article the author merely outlines the relevant computations: the limitations of a paper medium prohibit complete documentation including experimental data, parameter values and the author's programs. Consequently, the reader has painfully to re-implement the author's work before verifying and utilizing it ... The reader must spend valuable time merely rediscovering minutiae, which the author was unable to communicate conveniently [Schwab *et al.*, zitiert nach Gentleman *et al.* 2004].

Letztlich sind undokumentierte Ergebnisse nicht Wissenschaft, sondern nur Reklame für Wissenschaft:

> An article about computational science in a scientific publication is *not* the scholarship itself, it is merely *advertising* of the scholarship. The actual scholarship is the complete software development environment and that complete set of instructions that generated the figures [Buckheit und Donoho 1995, S. 59, nach Claerbout].

PISA-Vertreter verweisen gerne darauf, dass die Daten offen zugänglich seien. Das sollte in öffentlich geförderter Forschung eigentlich eine Selbstverständlichkeit sein, aber es ist nicht genug. So wie für die didaktische Auseinandersetzung mit PISA die Offenlegung der Aufgaben zu fordern ist, so für die Bewertung der Datenaufbereitung die Offenlegung der Software.

> Publication of the data from which articles are derived is becoming the norm [...]. This practice provides one of the components needed for reproducible research – access to the data. The other major component that is needed is access to the software and the explicit set of instructions or commands that were used to transform the data to provide the outputs on which the conclusions of the paper

> rest [...] It is easy to identify major publications in the most prestigious journals that provide sketchy or indecipherable characterizations of computational and inferential processes underlying basic conclusions. This problem could be eliminated if the data housed in public archives were accompanied by portable code and scripts that regenerate the article's figures and tables [Gentleman *et al.* (2004)].

Damit ist freilich nicht zu rechnen, solange Verantwortliche in den Teilnehmerstaaten dafür keinen Bedarf sehen oder nicht einmal den Unterschied zwischen Offenlegung und Kommerzialisierung eines Programms verstehen (Anh. D).

3.14 Umrechnung zwischen Prozenten und Punkten

Nach diesem Überblick über die Skalierung der PISA-Daten kann nun eine approximative Umrechnung zwischen prozententualen Lösungshäufigkeiten und PISA-Punkten angegeben werden. Gemäß Antwort- und Bevölkerungsmodell (13) wird auf eine Aufgabe i mit der Häufigkeit

$$\rho(k;\xi_i) = \int \mathrm{d}\theta\, A_i(k,\xi,\theta)\, \mathcal{N}(\theta;\, 0,\, \delta) \tag{28}$$

eine Antwort der Kategorie $k \in \mathcal{K}_i$ gegeben. In Verbindung mit den Transformationen (24) und (25) erwartet man für die Häufigkeit einer richtigen Antwort bei dichotomen Aufgaben in externen Einheiten

$$\rho(\xi^{\mathrm{P}}) = \int \mathrm{d}\theta^{\mathrm{P}}\, A^{\mathrm{P}}_{\mathrm{Rasch}}(\text{richtig}, \xi^{\mathrm{P}}, \theta^{\mathrm{P}})\, \mathcal{N}(\theta^{\mathrm{P}};\, 510{,}47,\, 100)\,. \tag{29}$$

Wegen der ungenügenden Reproduzierbarkeit der offiziellen Auswertung ist es wichtig, eine solche theoretisch hergeleitete Beziehung an veröffentlichten Daten zu überprüfen. Deshalb ist (29) in Abb. 3 eingezeichnet. Die gestrichelte Kurve gibt den Zusammenhang zwischen ρ_i und ξ_i im Großen und Ganzen korrekt wieder; zugleich verdeutlicht sie die unverstandene, scheinbar zufällige Nichtmonotonie. Außerdem sind in der Abbildung die falsche, aber von (29) kaum unterscheidbare Rekonstruktion

$$\rho(\xi^{\mathrm{P}}) = \frac{1}{1+\exp(-(500-\xi^{\mathrm{P}})/100 - \ln(62/38))} \tag{30}$$

aus der ersten Fassung dieses Aufsatzes (W1, Gl. (2) und Abb. 4) [gepunktet] und die ebenfalls unzutreffende, auch empirisch danebenliegende Rekonstruktion aus der Replik von Prenzel und Walter (2006) [durchgezogen] eingezeichnet; diese Irrtümer sind in Anhang B näher erklärt.

Aus der Steigung der gestrichelten Kurve oder durch Differentiation von (30) findet man die Umrechnung zwischen Lösungshäufigkeiten und Aufgabenschwierigkeiten: vier Schwierigkeitspunkte machen maximal einen Prozentpunkt in der Lösungshäufigkeit aus.

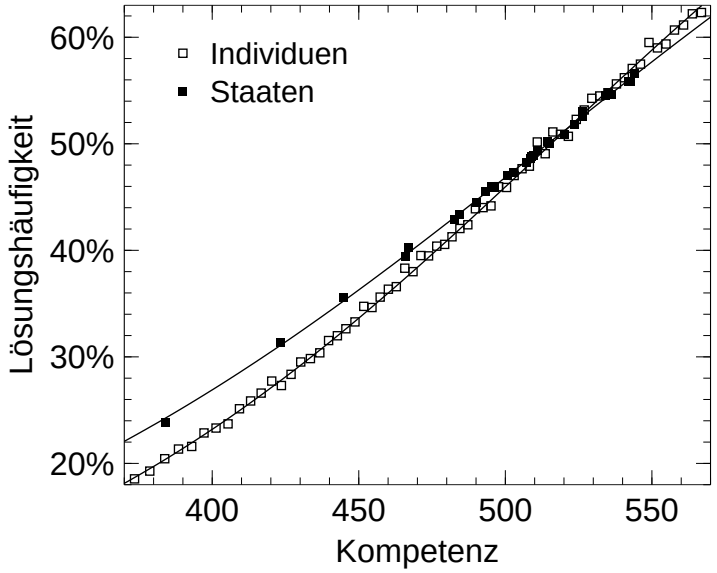

Abbildung 4: Lösungshäufigkeit, gemittelt über die 76 dichotomen Mathematikaufgaben, als Funktion der offiziellen Kompetenzwerte. Offene Symbole: alle OECD-Probanden in 100 Gruppen eingeteilt; angepasst mit einer Rasch-Funktion der Breite 97. Geschlossene Symbole: Mittelwerte für 30 OECD-Staaten; angepasst mit einer Rasch-Funktion der Breite 115.

Komplementär dazu wird in Abbildung 4 die Umrechnung zwischen Lösungshäufigkeiten und *Kompetenzwerten* untersucht. Deren Zusammenhang hängt von der empirischen Verteilung der 84 Aufgabenschwierigkeiten ab und lässt sich deshalb nicht einfach als ein zu (29) analoges Integral schreiben; er lässt sich jedoch gut durch eine Rasch-Funktion variabler Breite nähern. Wenn man OECD-weit die Probanden ihren plausiblen Leistungswerten nach in 1 %-Quantile einteilt, erhält man eine Rasch-Breite von 97; das heißt, auch vier Kompetenzpunkte entsprechen maximal ziemlich genau einem Prozentpunkt in der Lösungshäufigkeit.

Wenn man die Probanden hingegen nach Staaten zusammenfasst, ergibt sich durch die Mittelung über die breite Leistungsverteilung innerhalb der Staaten eine flachere Kurve mit einer Rasch-Breite von 115. Ein mittlere Kompetenzdifferenz von vier Punkten zwischen zwei Staaten bedeutet also einen Unterschied von weniger als einem Prozentpunkt in der Lösungshäufigkeit.

Der offiziellen Auswertung zufolge können 9 Punkte bereits als „signifikanter“ Leistungsunterschied zwischen zwei Staaten gelten. Staaten, die sich um nur 9 Punkte unterscheiden, können im OECD-Ranking um bis zu vier Plätze auseinander liegen. Auch der Unterschied von 10 Punkten zwischen den deutschen Ergebnissen in den Testgebieten Mathematik und Problemlösen ist ernst genommen und inhaltlich gedeutet worden (Prenzel *et al.* 2004a, im weiteren zitiert als D03a, S. 15).

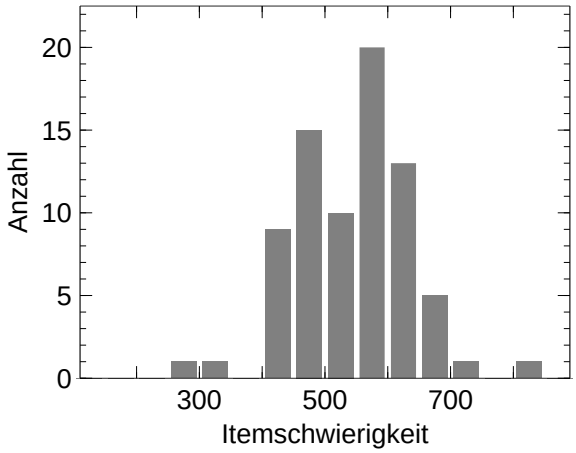

Abbildung 5: Schwierigkeitsverteilung der 76 dichotomen Mathematikaufgaben

Die obigen Abschätzungen zeigen, dass neun Punkte einer Differenz der Lösungshäufigkeiten von etwa 2 % entsprechen. Da im Mittel jedem Schüler knapp 26 Mathematikaufgaben gestellt wurden, entsprechen 9 Punkte ziemlich genau einer halben Aufgabe. Auf eine halbe Mathematikaufgabe entfallen 75 Sekunden Testzeit. Damit ist klar, dass die Gesamtergebnisse von PISA empfindlich von der Validität jeder einzelnen Testaufgabe abhängen. Auch ein von Land zu Land unterschiedlich strenger Umgang mit der Testzeit kann zu erheblichen Verzerrungen führen.

3.15 Verteilung der Aufgabenschwierigkeiten

Bei der Zusammenstellung von Testaufgaben muss man zwischen verschiedenen Anforderungen abwägen: Um Probanden verschiedenster Kompetenz mit gleicher Genauigkeit bewerten zu können, ist eine möglichst gleichmäßige, breite Verteilung der Aufgabenschwierigkeiten anzustreben. Um die Testmotivation über zwei Stunden und fünfzig Aufgaben hinweg aufrecht zu erhalten, wäre ein gewisses Übergewicht leichter Aufgaben sinnvoll. Um Störungen durch Probanden, die frühzeitig abgeben, zu vermeiden, darf der Test aber auch nicht zu leicht sein.

Wie Abbildung 5 zeigt, erfüllt der Mathematik-Test aus PISA 2003 nur die letztgenannte Anforderung: Er ist ziemlich schwer. Der Median der Schwierigkeitswerte der dichotomen Aufgaben liegt bei 552. Nur zwei Aufgaben liegen unter 400, aber zwanzig über 600.

Dieses Ungleichgewicht spiegelt sich auch in Abbildung 4 wider: Schüler mit einer Kompetenz von 500 Punkten lösen deutlich weniger als 50 % aller Aufgaben. Für einen erheblichen Teil aller Probanden dürfte PISA 2003 eine ziemlich frustrierende Erfahrung gewesen sein.

4 Was testen die einzelnen Aufgaben?

In diesem Teil wird die Antwortstatistik einzelner Testaufgaben untersucht. Es wird gezeigt, dass die Lösungshäufigkeiten als Funktion der den Probanden zugeschriebenen Kompetenzwerte durch die einparametrige Rasch-Funktion nicht adäquat modelliert werden; es gibt unterschiedliche Trennschärfen (4.2), kompliziertere Funktionsverläufe (4.3), Rateeffekte (4.4). Deshalb sind die vom Konsortium ermittelten Aufgabenschwierigkeiten auf zig Punkte ungenau (4.5). Bei der weiteren Analyse treten Dimensionen des Testgeschehens in den Vordergrund, die wenig mit Fachkompetenz zu tun haben, wie Vertrautheit mit dem Aufgabenformat (4.6), kultureller und sprachlicher Hintergrund (4.7 f.) sowie Ausdauer und Teststrategie (4.9).

4.1 Modelltests, Lösungsprofile

Die Antwortmodelle, auf denen die quantitative Auswertung von PISA beruht, beruhen ihrerseits auf einer ganzen Reihe von Annahmen: Alle Aufgaben eines Gebiets prüfen dieselbe Fähigkeit, ihre Trennschärfen sind ähnlich, Missverständnisse und Raten spielen keine Rolle, die Testleistung ist nicht zeitbegrenzt, und so weiter. Sind einzelne Annahmen verletzt, äußert sich das mit hoher Wahrscheinlichkeit in Abweichungen zwischen Lösungsstatistik und Modell.

Um solche Modellverletzungen festzustellen, steht eine ganze Batterie statistischer Tests zur Verfügung (z. B. Glas/Verhelst in Fischer/Molenaar 1995). In PISA scheinen nur zwei statistische Maße für die Modellgültigkeit verwendet worden zu sein: die Trennschärfe und ein unsäglich schlecht erklärter „infit“ (TR, S. 123). Bestimmte Akzeptanzgrenzen scheinen nicht vorgegeben worden zu sein. Die Kosequenz, bei Ungültigkeit des Rasch-Modells auf andere Modelle auszuweichen (Glas/Verhelst *loc. cit.*, S. 94), war wahrscheinlich schon durch die politische Zielvorgabe, Ranglisten zu liefern, ausgeschlossen.[37]

Es besteht also dringender Forschungsbedarf, die Gültigkeit der in PISA verwendeten Antwortmodelle zu prüfen. Wie die deutschen Mathematikdidaktiker (D. Lind *et al.* 2005, S. 83 f.) bestätigen, ist Modellvalidität zumindest auf der Ebene der Gesamtpopulation eine notwendige Voraussetzung für Gruppenvergleiche. Im folgenden soll die Validität des Raschmodells für eine Auswahl dichotomer PISA-Aufgaben geprüft werden. Dabei wird bewusst auf statistische Tests verzichtet, die auf die gedankenlose Verifikation von Nullhypothesen hinauslaufen und die für *jeden* empirischen Datensatz, wenn er nur umfangreich genug ist, „signifikante“ Modellverletzungen anzeigen (Hambleton *et al.* 1991, S. 53; allgemeiner Gill 1999, S. 657 f.).

[37] Köller (2006a, Punkt 5) zufolge sieht sich das Konsortium außer Stande, plausible Werte für Mehrparameter-Modelle zu bestimmen, weil dafür keine Software bereitsteht.

Sehr viel aussagekräftiger ist es, empirische Daten zusammen mit der Funktion, durch die sie modelliert werden, graphisch aufzutragen und die Übereinstimmung mit dem bloßen Auge zu beurteilen (Hambleton *et al.* 1991, S. 66; Andersen mit Berufung auf Rasch in Fischer/Molenaar 1995, S. 387; allgemeiner Meehl 1978, S. 825). Dazu werden die plausiblen Werte sämtlicher Probanden geordnet und in 20 oder 25 Gruppen gleichen Umfangs eingeteilt. Für jede Gruppe γ werden Mittelwerte der Kompetenz θ_γ und der relativen Lösungshäufkeiten $\rho_{i\gamma}$ berechnet.[38] Die Auftragung von $\rho_{i\gamma}$ über θ_γ soll im folgenden als *Lösungsprofil* einer Aufgabe i bezeichnet werden. Im Technischen Bericht wird ein einziges solches Profil mitgeteilt (*score curve*, S. 127), und zwar ein atypisch modellkonformes (mit Trennschärfe nahe am Sollwert 1/77,89).

4.2 Trennschärfe

Die Trennschärfe gibt an, wie effizient eine Aufgabe zwischen schwächeren und stärkeren Schülern unterscheidet. Gleichung (26) kann man so lesen, dass allen Mathematikaufgaben die Trennschärfe 1/77,89=0,01283 zugeschrieben wird: auf der Ebene des einzelnen Probanden bewirkt ein Kompetenzanstieg um 4 Punkte eine Erhöhung der Lösungswahrscheinlichkeit um 1,28 Prozentpunkte. Aus den in Kapitel 13 des Technischen Berichts angegebenen Skalentransformationen kann man ersehen, dass die vier Teiltests von PISA 2003 unterschiedliche Trennschärfen haben; beispielsweise hat der Lesetest die Trennschärfe 0,01100. Innerhalb eines jeden Testgebiets wird in der offiziellen Datenauswertung eine einheitliche Trennschärfe angenommen. Dafür gibt es keinen theoretischen Grund; es ist allenfalls denkbar, dass eine strenge Vorauswahl der Aufgaben zu einigermaßen ähnlichen Trennschärfen geführt hat.

Abbildung 6 zeigt, dass das nicht der Fall ist: sowohl unter den relativ leichten als auch unter den relativ schweren Mathematikaufgaben finden sich trennstarke („Exchange Rate Q1“, „Running Tracks Q2“) und trennschwache („View Room Q1“, „Carbon Dioxide Q3“). Die Trennschärfe ist also ein von der Schwierigkeit unabhängiger, quantitativ relevanter Aufgabenparameter. Das hat eine unmittelbare Auswirkung auf die Interpretierbarkeit numerischer Ergebnisse: Sobald man anerkennt, dass die Trennschärfe als ein zweiter Aufgabenparameter berücksichtigt werden muss, kann man den ersten Parameter, die Schwierigkeit,

[38] Um Ergebnisse möglichst unmittelbar mit denen der offiziellen Auswertung vergleichen zu können, passe ich mich an die Gewichtung der Probanden in der „international calibration“ an, das heißt, ich beziehe Großbritannien ein, gewichte alle 30 OECD-Staaten gleich, gewichte auch innerhalb der Staaten alle Probanden gleich, schließe Kurzhefte aus und werte „nicht erreichte“ Aufgaben als nicht gestellt. Hierdurch ergeben sich in Abbildungen und Fitergebnissen geringfügige Änderungen gegenüber W1.

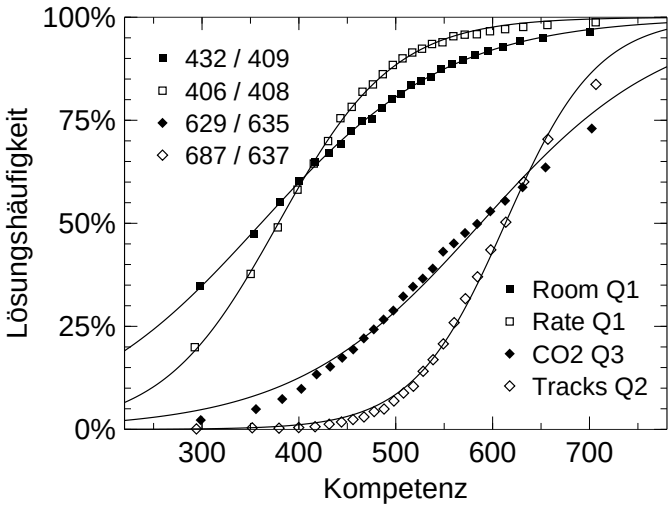

Abbildung 6: Lösungsprofile für zwei Paare von Mathematikaufgaben mit jeweils ungefähr gleicher Schwierigkeit, aber sehr unterschiedlicher Trennschärfe. In der Legende sind unten rechts Kurznamen der Aufgaben angegeben, oben links die Schwierigkeiten, und zwar zuerst laut Technischem Bericht und dann so, wie sie sich aus meiner Anpassung ergibt. Die durchgezogenen Kurven folgen dem zweiparametrigen Modell (31).

nicht mehr zur Feststellung einer eindeutigen Rangordnung der Aufgaben nutzen. Eine Änderung der Verankerung der nach außen kommunizierten Schwierigkeitsskala von 62 % auf einen anderen willkürlichen Wert würde sich nämlich bei trennstarken Aufgaben stärker als bei trennschwachen auf die Schwierigkeitsbewertung auswirken.

Abbildung 6 zeigt überdies, dass die 62 %-Festlegung in der offiziellen Auswertung nicht korrekt berücksichtigt wurde. Die Lösungsprofile der beiden Aufgabenpaare schneiden sich jeweils in der Nähe von 62 %. Es wäre deshalb zu erwarten, dass sich die Schwierigkeitswerte innerhalb der Paare „Room“ und „Rate“ sowie „CO2“ und „Tracks“ nur geringfügig unterscheiden. Tatsächlich aber liegen die offiziellen Schwierigkeitswerte (TR, S. 412f.) weit auseinander: um 26 bzw. 58 Punkte.[39]

Für eine adäquate Auswertung solcher Aufgaben ist es erforderlich, eine Trennschärfe D als zweiten Parameter in die Modellierung des Antwortverhaltens einzubeziehen. Diese Erweiterung des Rasch-Modells (26) zu einem zweiparametrigen logistischen Modell (2PL, gelegentlich als Birnbaum-Modell bezeichnet: Fischer in Fischer/Molenaar 1995, S. 19) lautet in externen Einheiten

$$A^{\mathrm{P}}_{2\mathrm{PL}}(\text{richtig}, \pi^{\mathrm{P}}, \theta^{\mathrm{P}}) = \frac{1}{1 + \exp\left[D^{\mathrm{P}}(\xi^{\mathrm{P}} - \theta^{\mathrm{P}}) - \ln(62/38)\right]}\,. \tag{31}$$

[39] Die Ursache liegt auf der Hand: Intern wurde das Rasch-Modell in der bei 50 % verankerten Form verwendet. Bei der Transformation (25) ist die unterschiedliche Trennschärfe offensichtlich nicht berücksichtigt worden, womit gegen die Modellierung (26)f. verstoßen wurde.

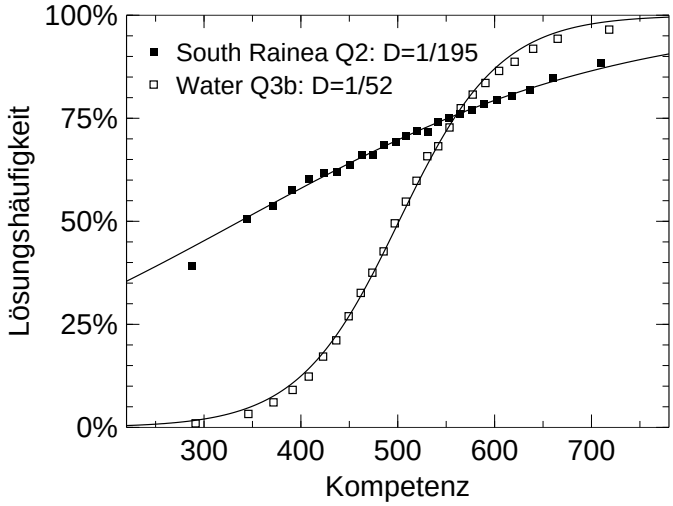

Abbildung 7: Lösungsprofile für zwei Naturwissenschaftsaufgaben mit extrem unterschiedlichen Trennschärfen. Die durchgezogenen Kurven folgen dem zweiparametrigen Modell (31); die schwächsten 4 % der Probanden wurden bei der Anpassung nicht berücksichtigt. Auch hier finde ich deutlich andere Aufgabenschwierigkeiten (Rainea 433, Water 527) als das Konsortium (466, 560).

Eine solche Verallgemeinerung ist mit dem oben (3.4 ff.) eingeführten Formalismus kompatibel; Gleichung (13) gilt unverändert. Somit könnte man die wahrscheinlichsten Werte der Aufgabenparameter $\underline{\pi}_i = (\xi_i, D_i)$ nach wie vor durch Maximierung der *Likelihood* $P(\underline{\underline{\kappa}}|\underline{\underline{\pi}}, \delta)$ bestimmen. Um eine solche Auswertung abzusichern, wäre es allerdings ratsam, zunächst die Skalierung mit dem einparametrigen Antwortmodell zu reproduzieren – was wegen der lückenhaften Dokumentation (3.2, 3.13) leider nicht möglich ist.

Stattdessen soll hier und in den folgenden Abschnitten ein einfacheres und anschauliches Verfahren zur Parameterschätzung angewandt werden, bei dem die Kompetenzwerte der Probanden als näherungsweise korrekt vorausgesetzt werden: Die jeweils betrachtete Antwortfunktion, hier (31), wird nach der Methode der kleinsten Quadrate an die empirischen Lösungsprofile angepasst. So sind schon die durchgezogenen Linien in Abb. 6 zustande gekommen. Sie beschreiben die empirischen Daten ziemlich gut; die Schätzwerte für die Aufgabenschwierigkeiten (408, 409 für die beiden leichten, 635, 637 für die beiden schweren Beispielaufgaben) passen, im Gegensatz zu den Ergebnissen des Konsortiums, perfekt zu der 62 %-Vorgabe.

Die Schätzwerte für die Trennschärfe liegen zwischen 1/97 („View Room“, „Carbon Dioxide“) und 1/48 („Running Tracks“). Noch weiter ist die Spanne im Naturwissenschaftstest (Abb. 7). Auch empirisch gibt es also nicht die geringste Rechtfertigung, alle Aufgaben eines Gebiets mit derselben Trennschärfe zu modellieren.

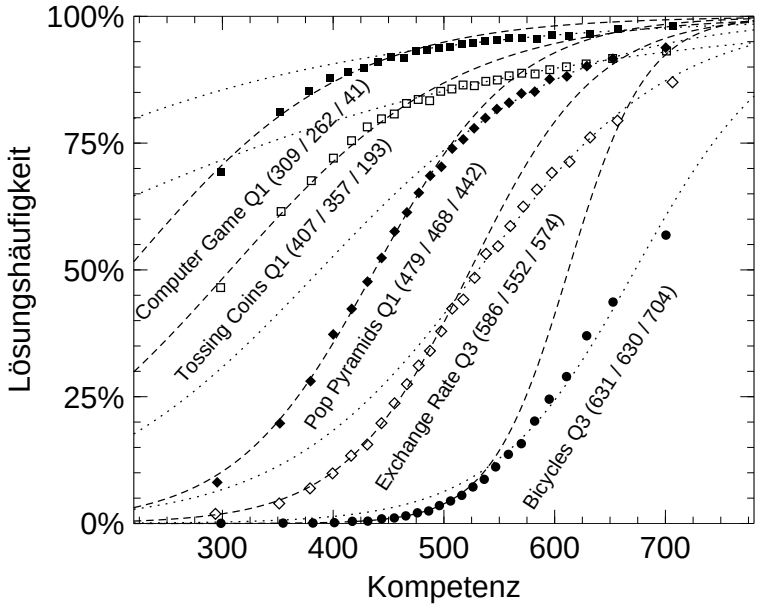

Abbildung 8: Zu fünf Lösungsprofilen wurde das zweiparametrigen Modells (31) je zweimal, getrennt für die leistungsschwächere (gestrichelt) und die stärkere (gepunktet) Hälfte der Schüler angepasst. In Klammern nach dem Aufgabennamen jeweils die ξ-Werte der offiziellen Auswertung, der gestrichelten und der gepunkteten Kurve.

4.3 Teilschritte oder alternative Lösungswege?

Etliche Aufgaben können auch mit zwei Parametern nicht angemessen modelliert werden. Abbildung 8 zeigt einige Beispiele, denen gemeinsam ist, dass das Lösungsprofil zunächst recht steil ansteigt, dann aber flacher verläuft als nach dem Zwei-Parameter-Modell (31) zu erwarten. Um die Unvereinbarkeit mit diesem Modell hervorzuheben, zeigt die Abbildung zu jeder Aufgabe zwei Anpassungen, die getrennt voneinander im unteren und im oberen Leistungsbereich vorgenommen wurden. Die Schwierigkeitswerte liegen um bis zu 220 Punkte auseinander, die Trennschärfen unterscheiden sich um Faktoren 1,5 bis 2,4. Die Schwierigkeitswerte der offiziellen Auswertung erweisen sich erneut als hochgradig unzuverlässig; sie sind nicht mit der 62 %-Verankerung kompatibel und liegen nicht einmal innerhalb der von den beiden Anpassungen eröffneten Spanne.

Natürlich gibt es keinen theoretischen Grund, die Schülerschaft scharf in zwei Hälften zu teilen. Numerisch stimmige und inhaltlich deutbare Anpassungen könnte man mit einem Vier-Parameter-Modell herstellen; ein solches Modell könnte eine „Reihenschaltung“ von zwei Lösungsschritten unterschiedlicher Schwierigkeit beschreiben. Genausogut könnte man an die Daten aber auch ein Fünf-Parameter-Modell anpassen, das eine „Parallelschaltung“ zweier verschiedener Lösungswege beschreibt (der fünfte Parameter ist die Häufigkeit, mit der einer der beiden Lösungswege gewählt wird). Noch realistischer wäre

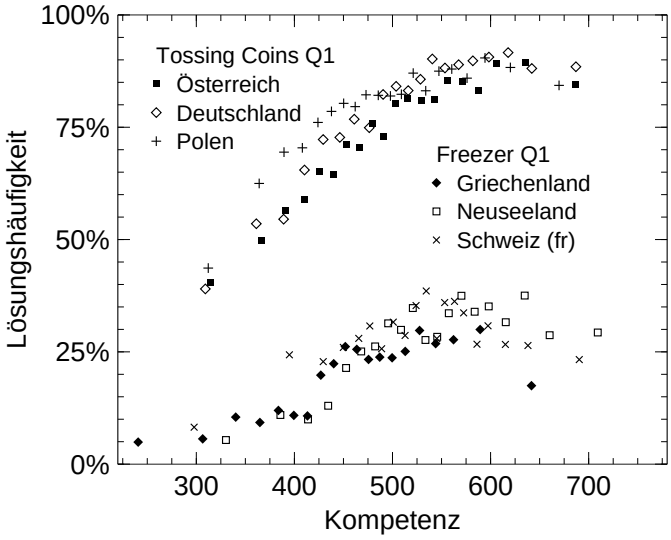

Abbildung 9: Nicht-monotone Lösungsprofile zweier Aufgaben in je drei Ländern/Regionen; ein Symbol entspricht hier jeweils 5 % der Schüler.

ein ganzes Netz von teils in Reihe, teils parallel geschalteten Teilschritten. Aber die in den Lösungsprofilen enthaltene Information ist zu unspezifisch, um eine nähere Klärung zugunsten eines bestimmten Modells herbeizuführen. Das hat unmittelbare Konsequenzen für die Deutung von Testergebnissen mittels Kompetenzstufen, denn verschiedene Modelle können zu weit auseinander liegenden Einschätzungen der Aufgabenschwierigkeit führen (vgl. Meyerhöfer 2004b, Bender 2005).

A priori ist nicht einmal sicher, dass die Lösungshäufigkeit eine monoton steigende Funktion der Schülerkompetenz sein muss. Die zweifelhafte inhaltliche Qualität einiger Aufgaben legt vielmehr die Vermutung nahe, dass ein Schüler, der sich fachlich auskennt und nicht mit der nächstliegenden, oberflächlichen Lösung begnügt, bei manchen Aufgaben benachteiligt sein könnte (Kießwetter 2002, Meyerhöfer 2005). Aufgaben, bei denen dies manifest in einzelnen Ländern der Fall ist, werden zwar als „psychometrisch nicht funktionierend“ aus der Auswertung herausgenommen und nicht näher dokumentiert. In PISA 2003 haben die Kontrollen jedoch bei mindestens zwei Aufgaben versagt: Abbildung 9 zeigt Aufgaben, bei denen in mindestens drei Regionen die ansonsten leistungsstärksten fünf, zehn oder fünfzehn Prozent der Schüler schlechter als die nächstschwächeren Perzentile abschneiden. „Freezer Q1“ wird auch in den übrigen Ländern von nur einem Drittel der Schüler im Sinne der Veranstalter gelöst – und das im leistungsstärksten Drittel nahezu unabhängig von der Kompetenz. Dass diese offensichtlich untaugliche Aufgabe in die Auswertung einbezogen wurde, zeigt die Unzuverlässigkeit der verwendeten Prozeduren.

4.4 Irgendetwas antworten

Die niederländischen Schüler ragen dadurch heraus, dass sie im Mittel weniger als 3,4 % aller Aufgaben unbeantwortet lassen. Mit beträchtlichem Abstand folgen zwischen 6,3 % und 8,0 % die fünf englischsprachigen Staaten, Finnland und Südkorea. In Deutschland, Österreich und der Schweiz liegt der Anteil fehlender Antworten zwischen 10,9 % und 11,3 %, in Dänemark über 14 %, in Italien über 19 %.

Noch konturierter wird das Bild, wenn man nur diejenigen Aufgaben betrachtet, die von besonders vielen Schülern übersprungen werden. Diese Aufgaben sind nicht im Multiple-Choice-Format, sind fast alle ursprünglich in Englisch oder Niederländisch und überwiegend von den Konsortialunternehmen CITO (Niederlande) und ACER (Australien) eingereicht worden, und sie sind fast alle unveröffentlicht.

Die acht auffälligsten Mathematikaufgaben wurden im Mittel von 10,8 % der niederländischen Schüler unbeantwortet gelassen. Es folgen mit 20,9 % die USA und bis 29,5 % die anderen englischsprachigen Staaten nebst Finnland, Südkorea und Island. Die weitere Spanne reicht über 44,6 % in Dänemark bis 55,6 % in Italien. Man vergleiche damit das Gesamtergebnis in Mathematik: 514 Punkte für Dänemark, 483 für die USA. Amerikanische Schüler haben demnach trotz schlechter fachlicher Voraussetzungen ausgesprochen geringe Hemmungen, auf schwierige oder abstruse Testaufgaben irgendeine Antwort zu geben – und niederländische Schüler sind den speziellen Stil des CITO wohl schon gewohnt, zumal die eine oder andere PISA-Aufgabe aus niederländischen Schulbüchern stammt (Jablonka 2006).

Dass manche Aufgaben sehr einfache, abgekürzte Lösungswege zulassen, zeigt sich auch in den Lösungsprofilen. Die beiden Multiple-Choice-Aufgaben in Abbildung 10 haben Plateaus im unteren Leistungsbereich: dort hängt die Lösungshäufigkeit nur schwach oder gar nicht von der Kompetenz ab. Die schwächste 4 %-Gruppe ist davon allerdings auszunehmen: ihre Lösungshäufigkeit liegt deutlich unter dem Plateau. Vielleicht liegt es an Zeitknappheit, vielleicht an Frustration, dass diese Schüler auf die Möglichkeit verzichtet haben, ihr Ergebnis durch Raten aufzubessern (vgl. Paris *et al.* 1991). Wenn man solche Dimensionen des Testgeschehens mit dem Hinweis ausblendet, die Fähigkeit, einen zweistündigen Test durchzustehen sei Teil dessen, was gemessen werden solle (Schulz 2006, Punkt 6), dann deutet die Modellverletzung durch die schwächsten 4 % darauf hin, dass weit unterdurchschnittliche Fähigkeitswerte häufiger vorkommen, als es per Normalverteilung vorausgesetzt wird (zu den Konsequenzen, die das für die Validität der Skalierung hat, siehe 3.5).

Auch wenn diese 4 %-Gruppe von der Auswertung ausgenommen wird, sind ein- oder zweiparametrigen Modelle hier natürlich inadäquat. Eine ungefähre Anpassung ist mit drei Parametern möglich, in Multiple-Choice-gewohnten

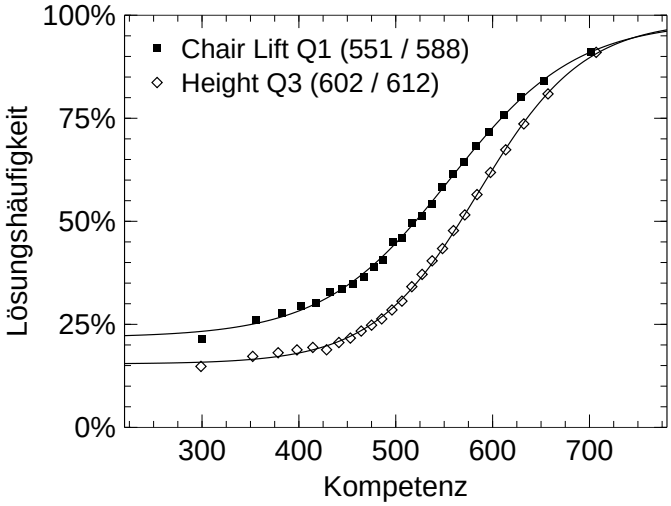

Abbildung 10: Lösungsprofile zweier Multiple-Choice-Aufgaben und Anpassung mit dem Vier-Parameter-Modell (32), das als alternativen Lösungsweg leistungsunabhängiges, qualifiziertes Raten annimmt. In Klammern die Schwierigkeitswerte aus der offiziellen und aus der hier gezeigten Auswertung.

Staaten auch durchaus üblich (Köller 2006a), inhaltlich jedoch nur begrenzt interpretierbar, denn wenn man Raten berücksichtigt, sollte man auch die Möglichkeit des Falschratens einbeziehen. Um der vollen Komplexität des Testgeschehens wenistens nahezukommen, ohne für jede Antwortalternative eigene Rasch-Parameter einzuführen (Andersen 1977), erscheint ein Vier-Parameter-Modell angemessen:

$$A^{\mathrm{P}}_{\mathrm{4Par}}(\mathrm{richtig}, \pi^{\mathrm{P}}, \theta^{\mathrm{P}}) = cr + \frac{1-c}{1+\exp(D^{\mathrm{P}}(\xi^{\mathrm{P}} - \xi^{\mathrm{P}}) - \ln(62/38))}\,. \quad (32)$$

Ein Bruchteil c aller Probanden bearbeitet die Aufgabe durch Raten; die übrigen Schüler wählen komplexere Lösungswege, die wie gehabt durch (31) approximiert werden. Es wird angenommen, dass *qualifiziert* (Meyerhöfer 2004a) geraten wird, so dass die Erfolgswahrscheinlichkeit r deutlich über $1/4$ liegen kann. Um die Zahl der Parameter nicht ausufern zu lassen, wird vereinfachend angenommen, dass r nicht von der Kompetenz abhängt. Abbildung 10 enthält eine Anpassung; die Schwierigkeitsparameter weichen auch hier deutlich von denen der offiziellen Auswertung ab.

4.5 Modellabhängigkeit der Aufgabenschwierigkeit

Die vorstehenden Beispiele zeigen deutlich, dass die Schwierigkeitsbewertung mancher Aufgaben um 30 oder mehr Punkte anders ausfallen kann, wenn anstelle des einparametrigen Rasch-Modells ein empirisch angemesseneres zwei- bis

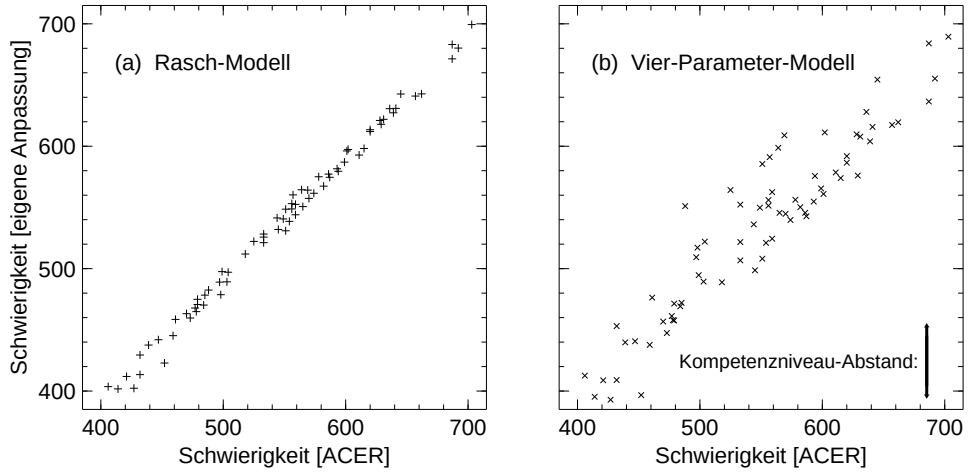

Abbildung 11: Schwierigkeitsparameter ξ_i für 73 von 76 dichotomen Mathematikaufgaben (die Werte für die übrigen drei Aufgaben liegen außerhalb des dargestellten Bereichs). Abszisse: Schwierigkeitswerte laut Technischem Bericht (TR, S. 412f.). Ordinaten: Schwierigkeitswerte aus eigener Anpassung (a) des Rasch-Modells (26) und (b) des Vier-Parameter-Modells (32) an Lösungsprofile wie die in Abb. 6ff. gezeigten.

vierparametriges Modell verwendet wird. Abbildung 11 zeigt die Streuung der Schwierigkeits-Schätzwerte für die Gesamtheit der dichotomen Mathematikaufgaben. Graph (a) validiert die hier gewählte Methode der Parameterschätzung, indem Aufgabenschwierigkeiten aus Anpassungen des eindimensionalen Rasch-Modells mit denen aus dem Technischen Bericht verglichen werden ($r = 0{,}997$ trotz mittlerer Differenz $-10{,}2$; Standardabweichung der Differenz 7,2; diese Abweichungen sind von ähnlicher Größenordnung wie die oben konstatierte intrinsische Inkonsistenz der vom Konsortium mitgeteilten Daten).

Im Vergleich zeigt dann Graph (b), wie sehr sich die Schwierigkeitsbewertungen bei Umstellung auf das Vier-Parameter-Modell ändern ($r = 0,948$, mittlere Differenz $-16{,}9$, Standardabweichung der Differenz 29,1). Diese Verzerrungen, in denen verkettete oder alternative Lösungsschritte (4.3) noch gar nicht berücksichtigt sind, sind mit der Breite der „Kompetenzstufen“ von 62 Punkten (LTW, S. 48) zu vergleichen: Eine beträchtliche Anzahl von Aufgaben würde bei Umstellung des Auswerteverfahrens einer anderen Kompetenzstufe zugeordnet.

Diese Schlussfolgerung hat Köller (2006a, Punkt 5) unter Berufung auf eine andere Studie und ungenannte Autoritäten bestritten. Ich analysiere seine Argumentation in Anhang D.

4.6 Multiple Choice: Mehrfachantworten

42 der 165 Aufgaben von PISA 2003 sind Auswahlfragen mit jeweils vier oder fünf Antwortalternativen, von denen genau eine als richtig gilt. In einigen Staa-

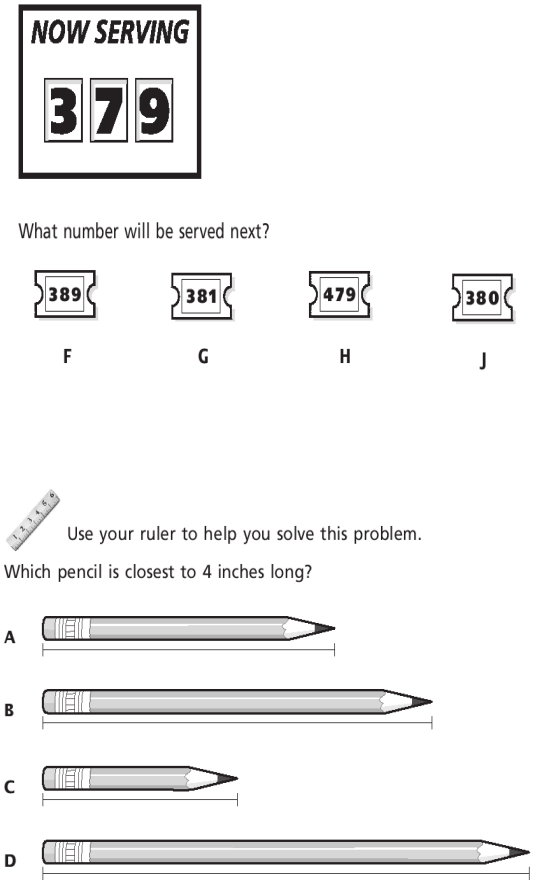

Abbildung 12: Zwei typische Aufgaben aus dem State Assessment für das 3. Schuljahr in Mathematik im Staat New York (NYSED 2006, mit freundlicher Genehmigung; Beispiel 2 nicht im Originalmaßstab). Sämtliche Aufgaben dieser Prüfung sind im Multiple-Choice-Format; immer ist genau eine von vier Antwortalternativen richtig.

ten, namentlich den USA, ist das das von der Grundschule bis ins Berufsleben vorherrschende Prüfungsformat.

Abbildung 12 zeigt zwei amerikanische Mathematikaufgaben für das 3. Schuljahr. Noch interessanter sind jedoch amerikanische Englischaufgaben. Beispielsweise wird zu einem Lesetext (NYSED 2005), der vom Winterschlaf eines Bären handelt, nicht nur *inhaltlich* gefragt

> According to the article, what do bears like to eat:
>
> (A) grass
> (B) honey
> (C) leaves and twigs
> (D) berries and nuts,

sondern es wird auch die Metaebene erklommen:

> The author **most likely** wrote this article to

(A) give readers information about a bear cub
(B) tell readers a funny story about a bear cub
(C) explain how bears survive cold winters
(D) describe what food bears like to eat.

Genau so sind die Leseaufgaben in PISA aufgebaut. In der Aufgabe „Graffiti“ wird mit einer expliziten Erklärung versucht, denjenigen Schülern nachzuhelfen, die diese Art Prüfungsfragen nicht von klein an gewohnt sind:

> We can talk about **what** a letter says (its content). We can talk about **the way** a letter is written (its style). Regardless of which letter you agree with, in your opinion, which do you think is the better letter? Explain your answer by referring to **the way** one or both letters are written [Kirsch *et al.* 2002, S. 52].[40]

Schwächen in der Konstruktion solcher standardisierter Testaufgaben erlauben immer wieder einmal abgekürzte Lösungswege. Die Fähigkeit, versteckte Hinweise (*secondary cues*) in einem Multiple-Choice-Test zu nutzen, wird seit über vierzig Jahren als „testwiseness“ diskutiert.[41] Für den deutschen Sprachraum hat Meyerhöfer (2005) das Wort „Testfähigkeit“ geprägt.

> Testwiseness has been a source of considerable concern to teachers – and amusement to students. The notion that it may be possible for a student to outwit a standardized test and perform well despite a significant lack of content-specific knowledge runs counter to principles of effective assessment.
> Students who are testwise are able to look for errors in the construction of test items, particularly in multiple-choice questions. Students who are able to outwit a test receive scores that are not valid, and not predictive of their current knowledge and skills or future abilities. It is important to differentiate between testwiseness and educated guessing. Testwiseness is based on little or no content knowledge and is merely an attempt to select the correct answer based on errors in test construction. In contrast, making educated guesses, requires the student to have some measure of content knowledge, enough at least to rule out some plausible distractors, reducing the number of possible answers from which a guess may be made [Mahamed *et al.* 2006].

[40] Die französische Version ist weniger klar: „En faisant abstraction de votre opinion, qui a ècrit la meilleure lettre, d'après vous? Justifiez votre rèponse en vous rèfèrant **à la façon** dont la lettre choisie est ècrite (ou à la façon dont sont ècrites les deux lettres).“ Die Probanden sollen also eine Meinung äußern („d'après vous“), nachdem sie von ihrer eigenen Meinung abstrahiert haben („On y pourrait perdre son latin“, Romainville 2002). Die Wiedergabe von „one or both“ ist überaus umständlich. Die deutsche Übersetzung ignoriert hier wie anderswo das gleichberechtigte französische Original und folgt eng der englischen Vorlage, benötigt aber über 30 % mehr Buchstaben als diese.

[41] Mahamed *et al.* (2006) schreiben den Begriff Gibb (1964, nicht geprüft) zu; eine andere Quelle (Millman *et al.* 1965) nennt noch frühere Vorgänger und fasst den Begriff bereits weiter, schließt zum Beispiel auch die Zeiteinteilung ein. Einen parodistischen Test, der ausschließlich Testfähigkeit testet, hat die New Yorker Schulbehörde auf ihren Webseiten versteckt (NYSED o. J.).

Die Frage liegt nahe, ob sich die unterschiedliche Vertrautheit mit dem Multiple-Choice-Format im internationalen Vergleich statistisch nachweisen lässt. Diese Frage ist schon für TIMSS gestellt worden und zumindest im Vergleich skandinavischer Staaten mit den USA negativ beantwortet worden (Lie *et al.* 1997, zitiert nach Olsen *et al.* 2001). Die PISA-Daten legen sogar nahe, dass Multiple-Choice-Aufgaben in Dänemark und Island besonders gut laufen; ein direkter Vergleich mit den USA wird außerdem durch die stark unterschiedliche Gesamtleistung erschwert. Wenn man hingegen das Gefälle in der Lösungshäufigkeit zwischen Multiple-Choice-Aufgaben und offenen Aufgaben zwischen den USA (58 % bzw. 41 %) und Irland (59 %, 47 %), Norwegen (58 %, 44 %), Ungarn oder der Slowakei (beide 59 %, 44 %) vergleicht, findet man die Erwartung, die amerikanischen Probanden könnten bei Auswahlaufgaben einen messbaren Vorteil haben, ansatzweise bestätigt. Der Unterschied von 2 bis 5 Prozentpunkten in der Lösungshäufigkeit entspricht 8 bis 20 PISA-Punkten, was eine erhebliche Verzerrung des internationalen Leistungsvergleichs möglich erscheinen lässt.

Einen eindeutigen Beleg für die unterschiedliche Vertrautheit mit dem Multiple-Choice-Format habe ich schließlich an anderer Stelle gefunden: nicht im prozentualen Anteil *richtiger* Lösungen, sondern in den unterschiedlichen Codes für *falsche* Antworten. In Staaten, in denen regelmäßig Multiple-Choice-Aufgaben eingesetzt werden, wissen die Studenten, dass stets nur *eine* Antwort korrekt ist . Häufige Mehrfachantworten (die in PISA mit einem eigenen Code gekennzeichnet und letztlich als falsch gewertet werden), sind ein starkes Indiz für mangelnde Vertrautheit mit dem Aufgabenformat.

Bei der unveröffentlichten Leseaufgabe „Optician Q1“ haben 10,5 % aller antwortenden österreichischen Schüler mehr als eine Alternative angekreuzt (Frankreich 8,6 %, Deutschland 8,1 %), während dieser Anteil in Australien, Island, Japan, Kanada, Südkorea, Mexiko, Neuseeland, den Niederlanden und den USA zwischen 0,0 % und 0,2 % liegt. Bei „Daylight Q1“[42] wurden die meisten Mehrfachantworten in Luxemburg (9,3 %), Österreich (8,5 %) und Deutschland (7,5 %) gegeben. Insgesamt haben in Deutschland, Luxemburg, Österreich bei 11, 12 bzw. 13 Aufgaben mehr als 4 % der Schüler eine Mehrfachantwort gegeben. Singulär ist die unveröffentlichte Problemlöse-Aufgabe „Cinema Outing Q2“, bei der selbst in den Niederlanden, Neuseeland, Australien über 10 % und in Katalonien volle 30 % mehr als eine Alternative für zutreffend gehalten haben. Dass auch diese offenkundig missratene Aufgabe in die offizielle Auswertung einbezogen wurde, zeigt erneut, wie unempfindlich die Kontrollprozeduren des Konsortiums sind. Dass das Aufgabenformat als bekannt vorausgesetzt wird,[43] zeigt den begrenzten kulturellen Horizont der Testentwickler.

[42] Dies ist eine der wenigen veröffentlichten Naturwissenschaftsaufgaben. Die vier Antwortalternativen sind *alle* falsch (Bender 2006).

[43] Das Testleiter-Handbuch (OECD 2003b) schreibt Wort für Wort vor, wie zu Beginn der Testsitzung einige Beispielaufgaben vorgestellt werden. Darunter ist auch eine Multiple-

Der hohe Anteil von Mehrfachantworten beweist, dass PISA-Ergebnisse in erheblichem Maße durch die unterschiedliche Vertrautheit mit dem Aufgabenformat verzerrt sind. Dass sich dieser Unterschied nur relativ schwach in der Abhängigkeit der Lösungshäufigkeiten vom Aufgabenformat äußert, mag daran liegen, dass *Testfähigkeit* auch bei den anderen Aufgabenformaten hilfreich ist. Der Stil der Leseaufgaben mit dem charakteristischen Wechsel zwischen inhaltlicher, sprachlicher und semiotischer Perspektive ist zum Beispiel sehr spezifisch für eine bestimmte Testkultur, ohne an ein bestimmtes Antwortformat gebunden zu sein. Und die Unkenntnis der Multiple-Choice-Grundregel verzerrt den Test weit über die betroffenen Aufgaben hinaus, denn sie bewirkt einen erheblichen Zeitverlust: es ist viel aufwändiger, vier oder fünf Antworten jeweils auf zutreffend oder nicht zutreffend zu prüfen, als eine einzige Alternative auszuwählen.

4.7 Weltwissen statt Leseverständnis?

„Optician Q1“ ist eine Auswahlaufgabe aus dem Testgebiet Leseverständnis. Tabelle 3 zeigt, mit welcher relativen Häufigkeit die Schüler zweier Staaten die vier Antwortalternativen gewählt haben. In beiden Staaten hat knapp die Hälfte der Schüler die als korrekt gewertete Antwort gegeben. Die übrigen Schüler haben im wesentlichen zwei andere Alternativen angekreuzt, und zwar mit nahezu spiegelbildlichen Häufigkeiten: in der Slowakei im Verhältnis 18 : 33, in Schweden im Verhältnis 37 : 14. Das heißt, bei formal beinahe identischer Testleistung unterscheiden sich die Präferenzen für die am häufigsten gewählten Distraktoren um rund 20 Prozentpunkte.

Ähnliche Verwerfungen finden sich bei einer ganzen Reihe anderer Aufgaben. Solange diese Aufgaben unveröffentlicht bleiben, ist eine Ursachenforschung kaum möglich. Ich zitiere deshalb aus dem Gedächtnis eine Analyse der veröffentlichten Aufgabe „Flu“ aus PISA 2000, die ich nicht weiterverfolgt und nicht näher dokumentiert habe. Als Textgrundlage sollte ein Firmenrundschreiben gelesen werden, das für eine Grippeschutzimpfung wirbt. In einer Multiple-Choice-Aufgabe sollten die Schüler dann angeben, wie das Rundschreiben die Schutzimpfung in Beziehung zu körperlicher Ertüchtigung und gesunder Ernährung setzt. Die korrekte Antwort entsprach einer Mittelposition; die Distraktoren gaben die Möglichkeit, der Schutzimpfung zuviel oder zuwenig zuzutrauen.

Choice-Aufgabe, bei der sich aber die Frage, ob mehr als eine Antwort richtig sein kann, nicht stellt: „Wo fanden 1972 die Olympischen Spiele statt?“ Es folgt die Instruktion „If you are not sure about the answer to a question, circle the answer that you think is best and continue with the next question on the test.“ Das kann man nicht ernsthaft als einen Hinweis werten, dass bei Multiple-Choice-Aufgaben *immer* nur eine Antwort richtig sein kann. Überdies zeugt es von wenig pädagogischer Erfahrung, eine so wichtige Regel mit einer einmaligen Ansage vermitteln zu wollen.

Tabelle 3: Relative Häufigkeit der vier Antwortalternativen der Multiple-Choice-Aufgabe „Optician Q1“. Die als korrekt gewertete Alternative „2“ wurde in den beiden aufgelisteten Staaten fast gleich häufig gewählt. Die Präferenzen für die Alternativen „3“ und „4“ unterscheiden sich hingegen um fast 20 Prozentpunkte.

	1	2	3	4
Slowakei	3,1 %	46,1 %	17,5 %	33,3 %
Schweden	3,1 %	46,2 %	37,0 %	13,7 %

Aus der Häufigkeitsverteilung der falschen Antworten ließ sich klar ablesen, dass französische Schüler die Impfung, deutsche Schüler gesunde Lebensführung für das wirksamere Mittel halten.

Oberflächliche Kenntnis der respektiven nationalen Mentalitäten hätte genügt, um dieses Ergebnis vorherzusagen. Aufgrund dieses Beispiels ist zu vermuten, dass ein erheblicher Teil der Schüler manche Aufgaben allein auf Grundlage allgemeinen Weltwissens beantwortet, ohne sich im geringsten auf den vorgelegten Text zu beziehen – was in Anbetracht des enormen Zeitdrucks, unter dem die Testung stattfindet, sogar eine vernünftige Strategie sein dürfte (vgl. Ruddock *et al.* 2006). Dann aber ist zu vermuten, dass auch die Häufigkeit *richtiger* Antworten nicht allein das Leseverständnis, sondern ebenso sehr Teststrategie und Weltwissen wiederspiegelt („the specifity of familiarity and content that makes comparability so problematic“, Goldstein 2004).

4.8 Sprachgruppen

Nach dem bis hierhin Gesagten ist klar, dass Voraussetzungen, die nichts mit fachlicher Kompetenz zu tun haben, einzelne Aufgaben erschweren oder erleichtern, und dass diese Voraussetzungen von Staat zu Staat variieren. Möglicherweise spielen Sprachunterschiede dabei eine noch größere Rolle als politische Grenzen. Um das näher zu untersuchen, ist es nützlich, Testergebnisse aus mehrsprachigen Staaten nach Sprachen aufzuschlüsseln.

Leider ist die Sprache, in der der Test durchgeführt wurde, nur für Belgien und die Schweiz im internationalen Datensatz enthalten. Andere mehrsprachiger Staaten wie Kanada, Spanien, Finnland oder Luxemburg halten diese wichtige Variable aus unbekannten Gründen unter Verschluss, selbst dann, wenn sie in nationalen Auswertungen durchaus berücksichtigt wird (z. B. Brunell 2004). Anhand von Testaufgaben, die in einzelnen Sprachen wegen Übersetzungs- oder Druckfehlern als nicht auswertbar kodiert wurden, lässt sich in einigen Fällen jedoch die Testsprache erschließen, so 2003 für die meisten Südtiroler Schulen. Die Luxemburger Schüler, die sich mit fünfzehn Jahren im Übergang von der deutschen Mittelstufe zur französischen Oberstufe befinden, haben sich ganz überwiegend auf Deutsch testen lassen.

In Finnland werden rund 5 % der Schüler in schwedischsprachigen Schulen unterrichtet. Im Datensatz von PISA 2000 lassen sich diese Schulen aufgrund einer in der schwedischen Fassung unbrauchbaren Aufgabe eindeutig identifizieren. Im schwerpunktmäßig getesteten Gebiet Lesen erzielen die finnischsprachigen Finnen 548 Punkte, während die schwedischsprachigen Finnen mit 513 geringfügig schlechter abschneiden als die Schweden mit 516 – und das, obwohl die schwedischsprachige Minderheit in Finnland laut Berufsprestige-Index einen überdurchschnittlichen sozio-ökonomischen Status genießt.[44]

Das legt die Vermutung nahe, dass das überragende Testergebnis Finnlands zu einem gewissen Teil auf einer sprachlich besonders zugänglichen Aufgabenübersetzung beruht. Weniger deutlich ist die Lage in Belgien, wo der Testleistungsunterschied zwischen den beiden großen Sprachgemeinschaften sogar stärker ist als der Unterschied zwischen den je gleichsprachigen Nachbarstaaten Niederlande und Frankreich. Hier zeigt sich, ähnlich wie oben bei der Suche nach Auswirkungen des Aufgabenformats, dass nationale oder regionale Leistungsmittelwerte von zu vielen verschiedenen Voraussetzungen abhängen, um einzelne Einflüsse zweifelsfrei nachweisen zu können – genau aus diesem Grund verlieren sich ja auch PISA-fromme Interpretationen in spekulativer Beliebigkeit.

Eindeutige Hinweise auf die Auswirkung von Staatszugehörigkeit und Testsprache erhält man erst, wenn man, anstatt über alle Aufgaben eines Gebiets zu mitteln, die Lösungsstrategien einzelner Aufgaben untersucht. Für die folgende Analyse werden die Testergebnisse der einzelnen Strata (Staaten oder Sprachgruppen) durch 660-komponentige Vektoren dargestellt. Jeder solcher Vektor enthält die Lösungshäufigkeiten für alle 165 Aufgaben und jeweils für alle vier Testhefte, in denen die Aufgabe vorkommt. Ähnlichkeiten im Schülerverhalten werden dann als Korrelationskoeffizienten r dieser Vektoren bestimmt.

Innerhalb der OECD findet man Korrelationen zwischen 0,979 (Australien – Neuseeland) und 0,743 (Japan – Mexiko).[45] Hohe Korrelationen werden vor allem durch eine gemeinsame Testsprache begünstigt; die staatliche Zusammengehörigkeit ist demgegenüber einflusslos. Zum Beispiel zeigt die deutschsprachige Schweiz die stärksten Korrelationen mit Deutschland (0,959), Luxemburg (0,956) und Österreich (0,954). Erst hinter dem niederländischsprachigen Belgien (0,937), Dänemark (0,930), Südtirol (0,926) und neun weiteren Staaten oder Regionen folgt die französischsprachige Schweiz mit einem Korrelationskoeffizienten von nur 0,910.

[44] Brunell (2004) kommt für PISA 2003 allerdings zum entgegengesetzten Schluss: der Abstand zwischen den beiden Sprachgruppen *verringere* sich, wenn man den sozio-kulturellen Status herausrechne. Auch ist zu berücksichtigen, dass es in den schwedischsprachigen Gebieten eine erhebliche Minderheit finnisch- oder zweisprachiger Schüler gibt.

[45] Zum Vergleich: innerhalb der einzelnen Staaten liegt der Korrelationskoeffizient zwischen Jungen und Mädchen zwischen 0,963 (Kanada) und 0,920 (Türkei).

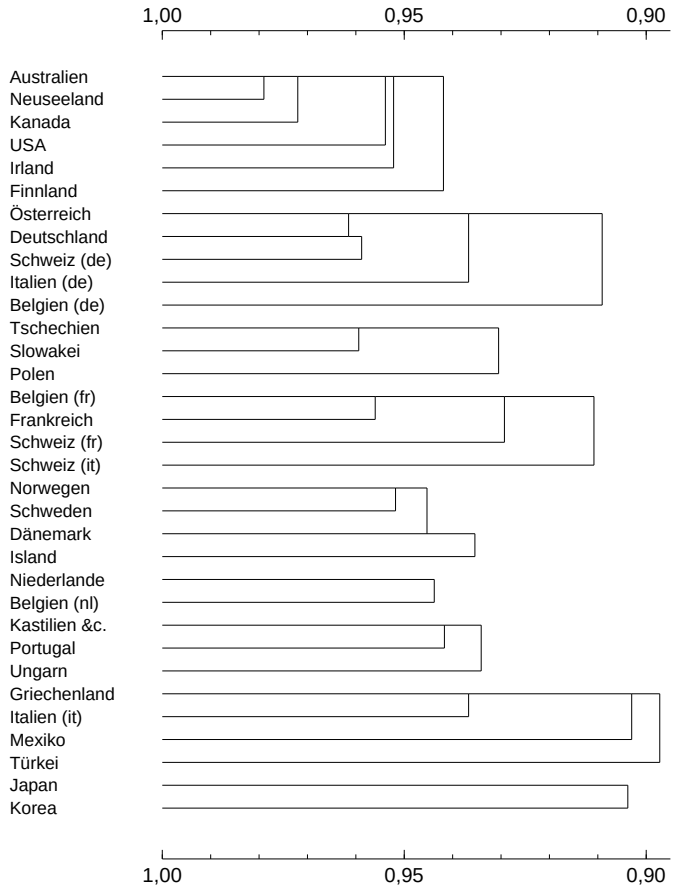

Abbildung 13: Dieses Diagramm veranschaulicht, in welchem Maße die Schüler verschiedener Staaten oder Sprachregionen dieselben Aufgaben mehr oder weniger erfolgreich lösen. Jede Region ist mit derjenigen verbunden, mit der ihr Lösungshäufigkeitsvektor die höchste Korrelation aufweist. Lesebeispiel: Den polnischen Ergebnissen stehen die tschechischen am nächsten (Korrelationskoeffizient r =0,931), den tschechischen aber die slowakischen (r =0,959).

Abbildung 13 visualisiert die Korrelationen durch Klammern, die jedes Stratum mit demjenigen verbinden, mit dem es am stärksten korreliert ist. Das zugehörige r lässt sich an der horizontalen Skala ablesen. Diese Auftragung führt zu Clustern von zwei bis sechs Regionen. Für die meisten Cluster sind die übereinstimmenden oder eng miteinander verwandten Testsprachen konstitutiv.[46]

Im wesentlichen übereinstimmende Cluster hat Rocher (2003) aus den Leseergebnissen von PISA 2000 abgeleitet. In PISA 2003 spielt das Testgebiet

[46]Der Klarheit halber sind die zweisprachigen Regionen Luxemburg, Baskenland und Katalonien ausgespart. Die zwei Fälle, in denen gleichsprachige Regionen keine bevorzugte Korrelation aufweisen (Italien/Schweiz, Mexiko/Spanien), erklären sich vermutlich mit dem großen Leistungsunterschied, der feinere Muster in den Lösungshäufigkeiten überdeckt; überdies wurden die Übersetzungen unabhängig voneinander erstellt (die italienischen sogar aus zwei verschiedenen Originalen, siehe Fußnote 49), und die Stichprobe der italienischen Schweiz ist sehr klein.

Lesen jedoch nur eine untergeordnete Rolle. Wenn man die Leseaufgaben aus der Auswertung herausnimmt, ändert sich Abb. 13 nur minimal; kein einziger Staat wechselt seine Clusterzugehörigkeit. Selbst wenn man sich nur auf die 34 Naturwissenschaftsaufgaben stützt, findet man noch ziemlich ähnliche Cluster, die sich auch bei Modifikationen der numerischen Prozeduren als robust erweisen (Olsen 2005a, 2005b).

Zum Teil mag das am „literacy"-Konzept liegen, welches mit sich bringt, dass nahezu jede PISA-Aufgabe erhebliche Leseanteile hat und die Grenzen zwischen den vier Testgebieten unscharf sind. Aber in den Grundzügen übereinstimmende Ähnlichkeitsbeziehungen zwischen Nationen hat Zabulionis (2001) schon aus TIMSS-Daten abgeleitet. Ohne weitergehende Untersuchungen ist es nicht möglich, aufzuschlüsseln, inwieweit diese Ähnlichkeiten auf gemeinsamen kulturellen, insbesondere curricularen Traditionen beruhen, und inwieweit auf sprachlichen Gemeinsamkeiten.[47]

Die systematisch mit Sprache und Kulturkreis korrelierte Variation der Aufgabenschwierigkeiten, die in den Clustern zum Ausdruck kommt, widerlegt die fundamentalen Modellannahme der offiziellen PISA-Auswertung, man könne Aufgabenschwierigkeiten durch einen einzigen, international einheitlichen Parameter beschreiben.

> With such a recognition, however, it becomes difficult to promote the simple country rankings which appear to be what is demanded by policymakers [Goldstein 2004].

Die Ranglisten, an denen die ganze Außenwirkung von PISA hängt, gründen sich auf nicht mehr als die vage Hoffnung, dass sich die je nach Staat und Sprache unterschiedlichen Schwierigkeiten in der Summe über alle Aufgaben irgendwie wegmitteln. Je weiter zwei Staaten sprachlich und kulturell voneinander entfernt sind, umso substanzloser ist diese Hoffnung, und umso stärker hängen sämtliche Vergleiche (der Testleistungen, der Standardabweichungen, der sozialen Gradienten und anderer Tertiärdaten) von der konkret getroffenen Aufgabenauswahl ab.

Die Aufgabenauswahl aber ist ein Ergebnis politischer Aushandlungsprozesse,[48] in denen der Bildungsbegriff der OECD operationalisiert wird. Bei der

[47] Bei einzelnen Aufgaben lässt sich der Einfluss kultureller Voraussetzungen quantitativ nachweisen: französische Schüler schneiden bei einem literarischen Text von Anouilh überdurchschnittlich gut ab (Rocher 2003), griechische Schüler bei Aesop.

[48] An einigen wenigen Literaturstellen scheinen diese Prozesse durch: Aushandlung zwischen Didaktikern und Psychometrikern (TR, S. 28); Aushandlung zwischen Ländervertretern (Prenzel 2004c). Wer etwas Gremienerfahrung hat, kann sich das ausmalen (vgl. Reagan-Cirincione/Rohrbaugh 1992, S. 182 f.). Ein hochrangiger französischer Regierungsvertreter berichtet, dass die Entscheidungsprozesse ziemlich opak und nicht unbedingt demokratisch seien, und dass es extrem schwierig sei, sich in der OECD Gehör zu verschaffen (Cyter-

Prüfung der intrinsischen Validität von PISA muss man diesen Bildungsbegriff und damit die grundsätzliche Ausrichtung der Testaufgaben als gegeben annehmen. Nichtsdestoweniger kann man fragen, ob nicht die kulturelle Herkunft und sprachliche Gestalt der konkret eingesetzten Aufgaben die *fairness* des Staatenvergleichs beeinträchtigen. Dieser Frage kann man sich nur nähern, indem man einzelne Quellen möglicher kultureller Verzerrung untersucht.

Eine solche Quelle ist die Herkunft der Testaufgaben (Bonnet 2002). Wie Ergebnisse aus PISA 2000 zeigen, sind Aufgaben in der Ursprungssprache typischerweise um 15 Punkte leichter als im internationalen Durchschnitt (Artelt/ Baumert 2004, S. 177). Wenn von den 129 Leseaufgaben die 18 auf Französisch eingereichten weggelassen würden, würde sich Frankreich alleine dadurch um 2,6 Punkte verschlechtern. Das ist fast soviel wie der Standardfehler des nationalen Mittelwerts (2,7). Teilt man den umfangreichen Lesetest aus PISA 2000 versuchsweise in einen Subtest mit original englischsprachigen Aufgaben und einen dazu komplementären Test auf, findet man für die englischsprachigen Staaten einen Vorteil zwischen 5 (USA) und 12 (Großbritannien) Punkten für den Test mit Aufgaben aus dem eigenen Sprachraum. Dass Artelt und Baumert fast alle unangenehmen Befunde für statistisch nicht signifikant erklären, ist irrelevant und ein offenkundiger Missbrauch der Methodik des Nullhypothesentests (vgl. Gill 1999, S. 661). Ihre Zusammenfassung, einem potentiellen *cultural bias* sei in PISA „durch eine möglichst multi-kulturelle Zusammensetzung von Testaufgaben begegnet“ worden, ist durch die in ihrem Aufsatz dargelegten Tatsachen nicht gedeckt.

Von der Frage der Herkunft zu unterscheiden ist das Übersetzungsproblem. In PISA werden eine englische und eine französische Aufgabenversion als gleichberechtigte Originale angesehen. Diese Originale sollten unabhängig voneinander in die Zielsprachen übersetzt und dann zusammengeführt werden; es gibt Anzeichen, dass es dabei drunter und drüber gegangen ist.[49]

---

mann in DESCO 2003, S. 22). Prenzel *et al.* (2007) versichern nichtsdestoweniger, dass die PISA-Aufgaben „gemäß *der* rationalen Testkonstruktion“ entwickelt werden (Hervorhebung des bestimmten Artikels von mir).

[49] In PISA 2000 wurde in über zehn Sprachräumen von der *empfohlenen* doppelten Übersetzung aus dem Englischen und Französischen abgewichen (Grisay in Adams/Wu 2002, S. 67). Diese Ausnahme galt nur dann als *akzeptabel*, wenn kein kompetenter Übersetzer aus dem Französischen aufgetrieben werden konnte (ebda., S. 61). Angeblich war das unter anderem in Portugal, Spanien, Griechenland und Südkorea der Fall. In Japan wurden nur Leseaufgaben aus dem Französischen übersetzt. Italien und die italienische Schweiz haben sich die Übersetzungen aufgeteilt, dann aber keine Zeit mehr zum Zusammenführen gehabt. In Dänemark, Finnland, Polen und den deutschsprachigen Ländern hat man angeblich eine doppelte Übersetzung aus dem Englischen gefertigt und dann *cross-checks* mit dem Französischen vorgenommen. (ebda., S. 67). Laut österreichischem Technischen Bericht (Haider 2001, IV 4.1) ist dagegen nur *eine* Übersetzung aus dem Englischen erfolgt. Diese wurde mit der Übersetzung aus dem Französischen unter extremem Zeitdruck auf einem Wochenendseminar zusammen-

Blum und Guèrin-Pace (2000, S. 113) berichten, dass allein die Umformulierung einer Frage („Quels taux ...?“) in eine Aufforderung („Énumèrez tous les taux ...“) die Lösungshäufigkeit um 31 Prozentpunkte anheben kann. Das lässt ahnen, welchen Spielraum die Übersetzer haben, Hilfestellungen zu geben oder zu verwirren (vgl. Freudenthal 1975, S. 172; Olsen *et al.* 2001, S. 411 ff.). Von den inhaltlichen *Verwerfungen*, die Meyerhöfer (2005) in seiner Analyse der „Bauernhöfe“-Aufgabe nachgewiesen hat, lassen sich mindestens zwei durch Übersetzungsfehler erklären: „Quader (rechtwinkliges Prisma)“ für „block (rectangular prism)“ und „Dachboden“ für „attic floor“ (in der österreichischen Fassung dagegen korrekt: „Boden des Dachgeschosses“).

Eine banale Folge von Übersetzungen ist, dass Texte dabei tendentiell länger werden. Herkunfts- und Übersetzungsproblem überlagern sich mit der Tatsache, dass Sprachen ohnehin unterschiedlich leicht lesbar sind. Am fassbarsten ist hier die unterschiedliche Textlänge. In PISA 2000 wurde für knapp sechzig Aufgaben ausgezählt, dass die Einleitungstexte im Französischen deutlich länger sind als im Englischen: sie haben durchschnittlich 12 % mehr Wörter und fast 19 % mehr Buchstaben.

Die differentielle Auswirkung allein der Wortanzahl auf die Lösungshäufigkeiten wurde leider nur anhand von Ergebnissen aus dem Feldtest untersucht und nur in grob zusammengefasster Weise veröffentlicht (Grisay in Adams/Wu 2002, S. 64 ff.). Man kann
immerhin herauslesen, dass allein dieser Effekt mit mehr als 1 % auf die Lösungshäufigkeit durchschlägt.

Zu dieser *differentiellen* Auswirkung auf einzelne Aufgabeneinheiten kommt noch der Zeitverlust im Gesamtverlauf des Tests. Selbst wenn man annimmt, dass die Lesegeschwindigkeit schwächer als direkt-proportional von der Textlänge abhängt, kommt man doch zu der Abschätzung, dass zehn bis zwanzig Prozent Unterschied in der Textlänge mehrere Prozent in der Lösungshäufigkeit, also zehn oder mehr PISA-Punkte ausmachen dürften.

> Ignoring the effects of multiple languages in a global society severely limits the validity of contemporary educational research [Sireci 1997].
>
> The common error is to be rather casual about the test adaptation process, and then interpret the score differences among the samples or populations as if they were real. This mindless disregard of test translation problems and of the need to validate instruments in the cultures where they are used has seriously undermined results from many cross-cultural studies [Hambleton 1994, zitiert nach Sireci].

---

geführt. Veröffentlichte Aufgabenbeispiele deuten darauf hin, dass das französische Original zumindest nicht ganz ernst genommen wurde. Beispiel: „a student“/„un èlève“ wurde mit „eine Schülerin“ übersetzt. Genau für solche Fälle, in denen eine Sprache klarer ist als die andere, war der ganze Aufwand der doppelten Übersetzung gedacht.

4.9 Leistungsabnahme und Zeitknappheit

Eine elementare Dimension des Testgeschehens, die in der offiziellen Auswertung erwähnt, aber nicht ernsthaft berücksichtigt wird, ist das Nachlassen der Schülerleistungen im Verlauf der Testsitzung. Dank des symmetrischen Testdesigns (Tab. 2) kann man diesen Effekt recht einfach quantifizieren: Der zweistündige kognitive Test ist in vier Blöcke gegliedert, die jeweils auf eine halbe Stunde ausgelegt sind. In jeder halben Stunde sind in der Summe über alle dreizehn Testhefte dieselben Aufgaben zu bearbeiten. Daher kann man die über einen Zeitblock gemittelten Lösungshäufigkeiten unmittelbar miteinander vergleichen. OECD-weit sinkt dieser Wert von 52,6 % im ersten auf 43,5 % im vierten Block (Tabelle 4, Spalte 4). Dieser Leistungsabfall entspricht gut 40 Kompetenzpunkten.

In W1 habe ich diese Abnahme der Lösungshäufigkeit als Ermüdung im weitesten Sinne, einschließlich Abnahme der Testmotivation, interpretiert. In den Reaktionen aus dem Konsortium (Schulz, Prenzel, Köller) wurden weder die Zahlen noch deren Erklärung in Frage gestellt. Mir selbst ist jedoch ein Zweifel gekommen. Den Probanden liegt während der gesamten Testsitzung das gesamte Testheft vor. Auf Rückfrage hat mir ACER bestätigt (Adams, Mail vom 17. 1. 2007), dass 2003 kein Einfluss darauf genommen wurde, in welcher Reihenfolge und mit welchem Tempo die Hefte bearbeitet wurden. Man kann also nicht davon ausgehen, dass auf jeden der vier Blöcke die gleiche Bearbeitungszeit von einer halben Stunde entfällt; zum Nachlassen der Lösungshäufigkeit, insbesondere vom dritten zum vierten Block, kann deshalb neben Ermüdung auch Zeitmangel beitragen.

Tabelle 4 liefert dafür Anhaltspunkte. Im ersten Block überspringen die Schüler 7,8 % aller Aufgaben. Im vierten Block lassen sie 19,7 % unbearbeitet; ein gutes Drittel davon (7,6 % aller Aufgaben) befindet sich am Ende des Testhefts und ist offiziell als „nicht erreicht“ kodiert worden; der Anteil übersprungener Aufgaben hat somit auf 12,1 % zugenommen. Wenn man sich nicht auf die Gesamtheit aller Aufgaben, sondern nur auf die bearbeiteten Aufgaben bezieht, dann fällt der Anteil richtiger Lösungen im Testverlauf um nur knapp drei Prozentpunkte ab. Der Rückgang der Lösungshäufigkeit im Testverlauf beruht somit nur zum geringeren Teil auf einer Zunahme falscher Antworten, sondern überwiegend darauf, dass zum Ende hin immer mehr Aufgaben übersprungen oder gar nicht erreicht werden.

Die Zeitknappheit, die sich in diesen Ergebnissen äußert, stellt eine theoretische Voraussetzung von PISA in Frage. In einem Lehrbuch über Testkonstruktion heißt es: Wenn nicht die Bearbeitungszeit für jede einzelne Aufgabe begrenzt werden kann, dann sollte die Testzeit

> so bemessen sein, dass in der Regel alle Personen bis zur letzten Testaufgabe vordringen. Nur in diesem Fall lassen sich die meisten Testmodelle auf die resul-

Tabelle 4: Im Testverlauf (vier halbstündige Blöcke) nimmt sowohl der Anteil bearbeiteter Aufgaben als auch der Anteil richtiger Lösungen ab (OECD-Staatenmittel ohne Großbritannien; ohne Sonderschulhefte; die Klassifizierung „nicht erreicht" beruht auf einem offiziellen Code für nicht bearbeitete Aufgaben am Ende eines Testhefts).

	nicht		richtige Lösungen pro		
	erreichte	bearbeitete	vorgelegte	erreichte	bearbeitete
	Aufgaben			Aufgaben	
Block 1	0,0 %	7,8 %	52,6 %	52,6 %	57,0 %
Block 2	0,2 %	9,3 %	51,4 %	51,5 %	56,7 %
Block 3	1,3 %	12,7 %	48,3 %	49,0 %	55,4 %
Block 4	7,6 %	19,7 %	43,5 %	47,1 %	54,2 %

tierenden Daten anwenden: Die unterschiedliche Anzahl von nicht bearbeiteten Aufgaben wirft rechnerische Probleme, vor allem aber auch Interpretationsprobleme auf [Rost 2004, S. 43].

Wenn ein Test unter Zeitdruck stattfindet, hängt das Ergebnis nicht nur von der Arbeitsgeschwindigkeit ab, sondern auch von der Stressresistenz, vom Zeitgefühl, von einem gewissen Mut zur Oberflächlichkeit und vom Überspringen schwieriger und zeitaufwändiger Aufgaben. Diese unspezifisch mitgetesteten Voraussetzungen können nicht einer *Kompetenz* in einem bestimmten Aufgabengebiet zugerechnet werden, sondern sind Teil einer generellen *Testfähigkeit*.

Die Stärke der bis hierhin am OECD-Mittel festgemachten Effekten ist international sehr unterschiedlich. Der Anteil nicht erreichter Aufgaben im vierten Block streut zwischen 1,0 % (Niederlande) und 25 % (Mexiko). Der Anteil übersprungener Aufgaben liegt anfänglich zwischen 2,5 % (Niederlande) und 11,5 % (Italien), zuletzt zwischen 4 % (Niederlande) und 21 % (Griechenland). Der Anteil richtiger Lösungen, bezogen auf die bearbeiteten Aufgaben, sinkt vom ersten zum vierten Block in der Schweiz und in Österreich um weniger als 2, in Griechenland und Island um über 5 Prozentpunkte. Bezogen auf die Gesamtheit aller Aufgaben, sinkt die Lösungshäufigkeit in Österreich um 4,7, in Griechenland um 17,6 Prozentpunkte.[50]

Welchen Einfluss diese Effekte auf den internationalen Leistungsvergleich haben, kann man größenordnungsmäßig abschätzen, indem man für jeden der vier halbstündigen Blöcke eine separate Rangliste der Lösungshäufigkeiten er-

[50]Es besteht eine gewisse Korrelation ($r = -0{,}746$) zwischen dem relativen Leistungsabfall und der anfänglichen Testleistung, aus der jedoch einige Staaten deutlich herausfallen. So haben Österreich und die Slowakei anfänglich die gleiche Lösungshäufigkeit 51 %; in der vierten halben Stunde liegt Österreich bei 46 %, die Slowakei bei 41 %.

stellt.[51] Die größten Veränderungen findet man im Mittelfeld. Österreich liegt in der ersten halben Stunde auf Platz 20, in der vierten halben Stunde auf Platz 12. Frankreich fällt hingegen von Platz 9 auf Platz 15. Irland und Ungarn liegen im ersten Block nahezu gleichauf, am Ende um acht Plätze auseinander. Eine Änderung einzelner Rahmenbedingungen (Testdauer, Bearbeitungsdauer pro Aufgabe) würde demnach genügen, um Ranking-Listen im Mittelfeld beliebig durcheinander zu schütteln; auch die Teststrategie (Raten, selektive Aufgabenbearbeitung) hat ganz erheblichen Einfluss auf das Gesamtergebnis.

Die von Land zu Land unterschiedlich schnelle Ermüdung kann verschiedenste Ursachen haben: Gewohnheit kürzerer oder längerer Schulstunden; Gewohnheit kürzerer oder längerer Leistungskontrollen; Länge und Ausgestaltung der Pause zwischen den beiden Teststunden; Frustrationstoleranz und Kritikfähigkeit gegenüber dem Test. Auch die zeitliche Nähe des Testtags zu Klassenarbeiten, Zeugnisterminen und Ferien dürfte die Ergebnisse beeinflussen. Vielleicht genügt ein unterschiedliches Osterdatum, um „signifikante“ Unterschiede zwischen verschiedenen Testjahrgängen vorzutäuschen.

Der Hinweis auf Ermüdungseffekte ist einer der wenigen Kritikpunkte aus W1, auf die das Konsortium inhaltlich eingegangen ist:

> The test length of two hours indeed causes fatigue effects and the consortium is also aware of the fact that these effects are different across countries. However, this does not invalidate the test results as the ability of doing a two-hour test [...] can be regarded as part of what is being tested [Schulz 2006, Punkt 6].

> Bemerkenswert sind schließlich die Hochrechnungen von Joachim Wuttke, wie die Testergebnisse bei unterschiedlichen Testzeiten ausgefallen wären. Würde man die international vorgegebene Testzeit von 120 Minuten bei dem vorhandenen Aufgabenmaterial verkürzen, ergäben sich selbstverständlich andere Leistungsverteilungen. Auch leistungsstarke Schüler[...] hätten dann keine Möglichkeit, anspruchsvollere Aufgaben erfolgreich zu bearbeiten [Prenzel 2006].

> „Wuttke hat außerdem bemängelt, die Testdauer habe Einfluss auf die Ergebnisse gehabt. Was sagen Sie zu diesem Vorwurf?“ – „Je länger Sie testen, desto präziser werden die Ergebnisse. Ich habe in meiner Stellungnahme versucht, das mit einem Fußballspiel zwischen einem Bundesligisten und einem Landesligisten zu vergleichen. Nach fünf Minuten wird's möglicherweise noch 0 : 0 stehen, nach fünfzehn Minuten vielleicht schon 3 : 0, nach neunzig Minuten wird es 8 oder 10 zu 0 stehen. Das heißt, je länger der Wettkampf oder die Testung stattfindet, desto klarer kommen auch die Unterschiede zutage“ [Interview Köller 2006b].

Keine dieser drei Stellungnahmen lässt sich auf die primäre, mathematische Implikation der Zeitabhängigkeit der Testleistung ein: dass eine solche Dimension,

[51] Hier bezogen auf die Gesamtheit der zu bearbeitenden Aufgaben und für 29 OECD-Staaten ohne Großbritannien.

die in der Item-Response-Modellierung des Schülerverhaltens nicht vorgesehen ist, systematische Messfehler verursacht, die in den offiziell mitgeteilten Standardfehlern nicht berücksichtigt sind.

Von Messfehlern zu reden, setzt freilich voraus, dass man den Gegenstand der Messung ernst nimmt. Insofern ist die Position von Schulz konsistent: Er bemüht sich überhaupt nicht, irgendeinen theoretischen Anspruch von PISA zu verteidigen. Er bestreitet nicht, dass unterschiedlich schnelle Ermüdung eine eigenständige Dimension des Testgeschehens ist. Sein Argument läuft darauf hinaus, dass PISA ein Aggregat verschiedener Fähigkeiten misst, zu denen auch Durchhaltevermögen gehört. Kurz gesagt misst PISA die Fähigkeit, zwei Stunden lang PISA-Aufgaben zu lösen. Auf diesem Niveau ist es weder möglich, noch erforderlich, PISA zu kritisieren.

Prenzel und Köller verteidigen die Dauer des Tests – als hätte ich eine Verkürzung der Testzeit vorgeschlagen. Das habe ich in W1 ebensowenig wie hier getan. Vielmehr argumentiere ich gegen eine verkürzte Interpretation von Testergebnissen; ich warne davor, eine Zahl, in die in völlig unkontrollierter Weise außer Wissen und Denkfähigkeit auch Durchhaltevermögen, Zeitmanagement und Teststrategie eingehen, als ein Maß für Kompetenz, „literacy“ oder was auch immer zu verwenden.

Warum Prenzel von einer Verkürzung der Testzeit bei Beibehaltung des vorhandenen Aufgabenmaterials redet, ist mir unerfindlich. Die Zeitknappheit verletzt schon jetzt die von Rost genannte Anforderung.

„Präzision“ scheint Köller allein im Sinne stochastischer Signifikanz zu verstehen. Aber selbst in dieser verengten Sicht, in der es keine Rolle spielt, wenn mit der größten numerischen Präzision der größte Mischmasch gemessen wird, ist sein Argument technisch falsch. Hier fällt der Vorwurf mangelnder Literaturkenntnis, den Köller gegen mich erhebt, auf ihn selbst zurück. Dass ein Test umso genauer sei, je länger er dauert, galt zwar lange als „Binsenweisheit“. Yousfi (2005) aber hat gezeigt, dass dieser „Mythos“ nur unter sehr engen Voraussetzungen fundiert ist. Sobald die stochastischen Unsicherheiten einzelner Aufgaben miteinander korreliert sind, kann die Verlängerung eines Tests zu Einbußen bei Gütekriterien führen.

5 Hintergrunddaten

In diesem Teil werden Aussagen über die Abhängigkeit der kognitiven Leistung von Hintergrundvariablen untersucht. Soziale Herkunft wird in PISA eindimensional durch einen Generalfaktor bewertet (5.1). Die Auswirkung dieses Faktors variiert sehr deutlich von Aufgabe zu Aufgabe (5.2). Auch Aussagen über den Zusammenhang zwischen Geschlecht und Testerfolg hängen kritisch von der Aufgabenauswahl ab (5.3).

5.1 Soziale Herkunft: der ESCS-Index

Mit dem *Questionnaire* wird eine Fülle von Angaben zum sozialen Hintergrund erhoben. Manche Auswertungen stellen auf spezifische Parameter ab; zum Beispiel hat in Deutschland das Geburtsland der Eltern besondere Aufmerksamkeit erfahren. In einem Großteil der offiziellen Auswertungen wird soziale Herkunft jedoch nicht näher aufgeschlüsselt, sondern anhand einer eindimensionalen Rangordnung bewertet.

In PISA 2000 wurde dieser Rangordnung allein der Beruf der Mutter oder des Vaters zugrunde gelegt.[52] Wenn ein Proband die Berufe beider Eltern angegeben hatte, wurde nur der höher eingestufte berücksichtigt. Die Zuordnung von Berufen zu Sozialprestigekennwerten (ISEI: „Standard International Socio-Economic Index of Occupational Status“) wurde einer Metastudie von Ganzeboom *et al.* (1992) entnommen, die 31 Berufsprestigeerhebungen aus 16 Staaten aufeinander skaliert und letztlich gemittelt haben. Die Primärdaten sind zum Teil über dreißig Jahre alt.[53] Der ISEI soll ausdrücklich nur das Sozialprestige *männlicher* Arbeitnehmer beschreiben. Nichtsdestoweniger wird er in PISA auch auf die Mütter der Probanden angewandt.

Tabelle 5 zeigt einen Auszug aus dem Katalog. Die Absurdität vieler Einstufungen und letztlich des ganzen Ansatzes ist offensichtlich. Die Korrelation zu einer durch Umfrage ermittelten deutschen Berufsprestige-Skala (Institut für Demoskopie Allensbach, 2005) beträgt lächerliche $r = 0{,}06$. Auch Mitglieder des Konsortiums äußern Vorbehalte; für die nationale Auswertung verwenden sie ein anderes Maß, das sie für „theoretisch besser fundiert“ und anschaulicher halten (Baumert *et al.* 2001, im weiteren zitiert als D00, S. 328).

[52] Gefragt wird nach dem zuletzt ausgeübten Beruf (DAM, S. 252). Eventuelle Arbeitslosigkeit (die erhebliche Korrelation mit Testleistungen der Kinder aufweist, Ebenrett *et al.* 2003) wird erfragt, aber nicht in die Sozialindizes eingerechnet.

[53] Ganzeboom *et al.* stützen sich auf eine Berufeliste der Weltarbeitsorganisation aus dem Jahr 1968 und auf Einzelstudien aus den Jahren 1968 bis 1982. Einige der aufgelisteten Berufe sind zwischenzeitlich ausgestorben (Aircraft Navigator, Telegraphist, Card- and Tape-Punching Machine Operator). Wie seither neu entstandene Berufe in PISA eingestuft werden, ist nicht dokumentiert.

Tabelle 5: Beispiele für die Einstufung von Berufen im „Standard International Socio-Economic Index of Occupational Status“ (ISEI, Ganzeboom *et al.* 1992). Welchen Status ein „Expert (not further specified)“ hat, verrate ich erst in Abschnitt 6.7.

ISEI	Berufe
90	Judge
88	Hospital Physician
83	Armed Forces Officer
79	Physicist
78	University Professor
73	Chemist, Member of Parliament
72	Political Scientist, Social Scientist, Psychologist, Large City Head
71	Statistician, Aircraft Pilot, High School Teacher, Middle School Teacher
69	Head of Large Firm, General Manager, Banker, Elementary School Teacher
65	Kindergarten Teacher, Teacher for the Blind
64	Large Shop Owner, Stock Broker, Computer Programmer, Dancer
61	Real Estate Agent, Astrologer
60	Sales Promotor, Department Head Provincial Government
58	Orthopedic Technician, Secretary, Soldier, Embalmer
56	Railway Station Master
55	Priest, Missionary, Faith Healer
54	Singer, Composer, Conductor, Clown
53	Tax Collector
52	Automobile Dealer
51	Dentist's Receptionist
49	Large Farmer, Restaurant Owner
48	Coffeeshop Operator, Air Traffic Controller
46	Cinema Projectionist
45	Meter Reader, Proofreader, Xerox Machine Operator
44	Aircraft Engine Mechanic, Airline Stewardess, Croupier
42	Head Nurse
39	Demolition Worker, Quality Inspector, Uncertified Nurse, Nurse Trainee
37	Elevator Operator, Shoe Shiner, Tobacco Factory Worker
36	Noodle Maker
35	Money Lender, Street Vendor, Telephone Solicitor, Museum Guard
34	Power-Reactor Operator
33	Brewer, Wine Maker, Ice-Cream Maker, Taxi Driver
32	Wine Waiter, Tree Surgeon, Oyster-Farm Worker, Trapper, Hunter, Whaler
30	Paving Machine Operator, Master Cook
29	Pig Farmer, Mushroom Grower, Musical Instrument Maker
28	Road Construction Worker, Soda Fountain Clerk
24	Clothes Washer, Chambermaid, Domestic Servant, Companion

ISEI	Berufe
20	Animal-Drawn Vehicle Driver, Poultry Farm Worker
18	Small Farmer
10	Cook's Helper, Apiary Worker, Picker, Gatherer

Bonnet (2002) berichtet, dass französische Regierungsstellen die OECD aus einem anderen Grund vor der Verwendung des ISEI gewarnt haben: Schülerauskünfte über Beruf und Ausbildung ihrer Eltern sind unzuverlässig. Als Reaktion wurde in den Ergebnisbericht (OECD 2001, S. 221) der Satz eingefügt: „In the case of France, questions remain about the reliability of students' responses regarding parental occupation and education". Die Einschränkung dieses Vorbehalts auf das Land, das ihn erhoben und belegt hat, ist eine Unverschämtheit.

In PISA 2003 wurde zur Beschreibung der sozialen Herkunft ein neuer Index gebildet (ESCS: *index of economic, social and cultural status*), der aus den folgenden drei Subindizes berechnet wird: (1) wie gehabt der ISEI des höher bewerteten Elternteils; (2) die aus dem höchsten Bildungsabschluss geschätzte Ausbildungsdauer des länger ausgebildeten Elternteils; (3) die Ausstattung des Haushalts mit bestimmten Gütern (DAM, S. 254f.). Begründet wird dieses Konstrukt mit genau zwei Sätzen:

> The rationale for using these three components is that socio-economic status is usually seen as based on education, occupational status and income. As no direct income measure is available from the PISA data, the existence of houshold items is used as an approximate measure of family wealth [TR, S. 318].

Die Ausgangsdaten beruhen auf teilweise zweifelhaften Schülerangaben und Kodierungen.[54] Einige Schüler haben den Hintergrundfragebogen nur unvollständig ausgefüllt; in manchen Sonderschulen wurde der Fragebogen gar nicht eingesetzt. Wenn zwei der drei Subindizes bestimmt werden konnten, wurde der dritte von ACER geschätzt; ansonsten wurde der ESCS-Index im internationalen Datensatz als „missing" kodiert. Nach Kanada (siehe 2.9) ist Deutschland das Land mit der höchsten Quote fehlender ESCS-Angaben (6,8 %, davon knapp die Hälfte in Sonderschulen), gefolgt von Großbritannien (4,7 %), der Tschechischen Republik (4,5 %) und den Niederlanden (3,9 %). Der Durchschnitt aller übrigen OECD-Staaten liegt bei 0,9 %.

[54] In Japan wurden nur 20, überall sonst 53 bis 63 verschiedene ISEI-Werte vergeben. Die Ausbildungsdauer, die zum höchsten Bildungsabschluss führt, beträgt dem Datensatz zufolge in Deutschland 17 Jahre, in der Schweiz 15 Jahre. In manchen Ländern ist angeblich ein beträchtlicher Teil der Eltern nie zur Schule gegangen (Ausbildungsdauer 0 Jahre; das ist *nicht* der Code für eine fehlende Angabe): Portugal 18,5 %, Mexiko 7,9 %, Luxemburg 4,8 %, Deutschland 4,6 %, Türkei 3,6 %, ... Griechenland 0,5 %, Österreich 0,2 %.

Auf dieser Datenbasis ist es kaum möglich, Deutschland in einen internationalen Vergleich einzubeziehen. Die vielpublizierten Aussagen zum in Deutschland besonders ausgeprägten Zusammenhang zwischen Testleistung und ESCS stammen denn auch nicht aus internationalen Berichten, sondern auf in Deutschland durchgeführten Auswertungen (D00, S. 381 ff.; D03b, S. 247 ff.). In diesen Auswertung wurden fehlende Angaben auch in den Fällen *imputiert* (qualifiziert erraten: Rubin 1987, Schafer 1997), in denen ACER (etwa nicht aus gutem Grund?) von einer Schätzung abgesehen hat.[55] Die Beschreibung dieser *missing data imputation* (D00, S. 334) ist geradezu exemplarisch mangelhaft.[56]

Passend zur theoriefreien Konstruktion des ESCS werden die Koeffizienten, mit denen die drei Subindizes gewichtet werden, rein empirisch aus einer Hauptkompentenanalyse bestimmt. Der ESCS ist also eine Art sozio-kultureller *Generalfaktor*. Er „erklärt" jedoch nur 53 % (Japan, Island) bis 68 % (Ungarn, Portugal) der Varianz seiner drei Eingangsvariablen, denn die sind nur mäßig miteinander korreliert; zum Beispiel beträgt die Korrelation zwischen elterlichen Ausbildungsjahren und Besitztümern nur 0,22 (Island) bis 0,52 (Portugal).[57] Das ist nicht ohne Ironie, denn auf der anderen Seite verwahrt sich das Konsortium heftig dagegen, die Ergebnisse aus den vier Gebieten des kognitiven Tests durch einen Generalfaktor zu beschreiben (dazu unten 6.1), obwohl die Korrelationen dort überaus deutlich sind (mehrheitlich über 0,8; Varianzaufklärung durch die erste Hauptkomponente zwischen 75 % in Griechenland und 92 % in den Niederlanden).

Die Beschreibung sozialer Herkunft durch *einen* Generalfaktor wird damit begründet, dass in sämtlichen Staaten nur *ein* Eigenwert der Kovarianzmatrix der drei Eingangsvariablen größer als eins ist (TR, Fußnote 9 zu S. 316). Das ist eine missbräuchliche Anwendung einer Faustregel (Guttman 1954, S. 154 f.), die nur bei einer wesentlich größeren Zahl von Ausgangsvariablen Sinn hat.

---

[55] Das wird nur im Bericht zu PISA 2000 näher beschrieben; Ergebnistabellen legen aber nahe, dass in PISA 2003 ähnlich vorgegangen wurde. Die geschätzten ESCS-Werte sind nicht im internationalen Datensatz enthalten und wohl auch nicht anderweitig veröffentlicht. Für eigenständige Analysen muss man Probanden mit fehlender ESCS-Angabe deshalb ausschließen, was den Mathematik-Duchschnittswert für Deutschland um 9,5 Punkte anhebt (Kanada, Tschechien, Niederlande 4 bis 6 Punkte, alle übrigen OECD-Staaten weniger als 2 Punkte) und die Vergleichbarkeit numerischer Ergebnisse entsprechend einschränkt.

[56] Mitgeteilt wird, dass man sich „leistungsfähiger Algorithmen" bedient, und dass das verwendete Computerprogramm einen schönen Namen hat. Erschließen kann man (wenn man die nicht erklärte Abkürzung „EM" mit „estimation–maximization" auflöst), dass die fehlenden Indexwerte als in einem probabilistischen Modell wahrscheinlichste Parameterwerte berechnet werden. Nicht mitgeteilt wird, worin dieses Modell besteht und welchen Input es verarbeitet.

[57] Im übrigen täuschen allein auf Korrelationen basierte Analysen leicht darüber hinweg, dass sich Unterschiede zwischen Staaten primär in einer absoluten Verschiebung der Skalen äußern. Beispielsweise ist ein Schweizer Haushalt bei gleichem Berufsprestige deutlich geringer mit kulturellen Statussymbolen ausgestattet als ein deutscher.

Ohnehin ist es reichlich absurd, die Dimensionalität eines Datensatzes mit einer Hauptkomponentenanalyse zu bestimmen, nachdem man diesen zunächst durch andere, willkürliche Auswertungsschritte auf drei Variable reduziert hat.

Sobald man weitere Variable wie zum Beispiel den Migrationshintergrund hier – und nicht in einer separaten Analyse – berücksichtigt, findet man, (1) dass es mehr als eine Hauptkomponente mit Eigenwert größer 1 gibt, soziale Herkunft also *nicht* statistisch adäquat durch nur *eine* Kennzahl beschrieben werden kann, (2) dass die erste Hauptkomponente einer solchen erweiterten Analyse deutlich vom ESCS abweicht, die Willkür in dessen Definition sich somit nicht herausmittelt, und (3) dass sich die Zusammensetzung der ersten paar Hauptkomponenten von Land zu Land stark unterscheidet, weshalb es unmöglich ist, die soziale Durchlässigkeit von Schulsystemen auf einer eindimensionalen Skala zu vergleichen. Mehrdimensionale Analysen werden dadurch begrenzt, dass Kinder aus benachteiligten Milieus sehr häufig das Questionnaire nur unvollständig bearbeiten (vgl. Hagemeister 2006).

5.2 Soziale Herkunft und Mathematikkompetenz

Zur weiteren Analyse des sozialen Phänomens PISA soll die Bewertung sozialer Herkunft durch den ESCS-Index unbeschadet aller Vorbehalte als gegeben angenommen werden. Außerdem soll hingenommen werden, dass das kognitive Testergebnis im Gegensatz zum sozialen Hintergrund nicht durch einen Generalfaktor ausgedrückt werden darf. Deshalb werden im folgenden, wie in den offiziellen Berichten, nur die Ergebnisse aus dem Schwerpunktgebiet Mathematik berücksichtigt.

Unter diesen Prämissen läuft die weitere Auswertung darauf hinaus, den statistischen Zusammenhang zwischen Mathematikkompetenz und ESCS zu quantifizieren. Das geschieht primär über lineare Regression. Die Steigung der Ausgleichsgeraden wird als sozialer *Gradient* bezeichnet. Der Gradient gibt also an, um wieviel Punkte sich die mittleren Testleistungen von zwei Subpopulationen unterscheiden, deren soziale Hintergründe um einen Indexpunkt auseinander liegen. Allerdings weichen die PISA-Berichte teilweise vom mathematischen Wortgebrauch ab und bezeichnen die Ausgleichsgerade selbst als „Gradient“.

Der deutsche Bericht verrät bemerkenswerte Unsicherheit über die Aussagekraft des sozialen Gradienten. Zunächst heißt es:

> Da es sich bei den sozialen Gradienten um unstandardisierte Koeffizienten handelt und diese von der Varianz abhängig sind, ist ein Vergleich zwischen zwei Messzeitpunkten problematisch [D03b, S. 247].

Nur zwei Absätze später hingegen:

> Durch den sozialen Gradienten wird der Zusammenhang von sozioökonomischem Hintergrund und dem erreichten Kompetenzniveau der Jugendlichen für

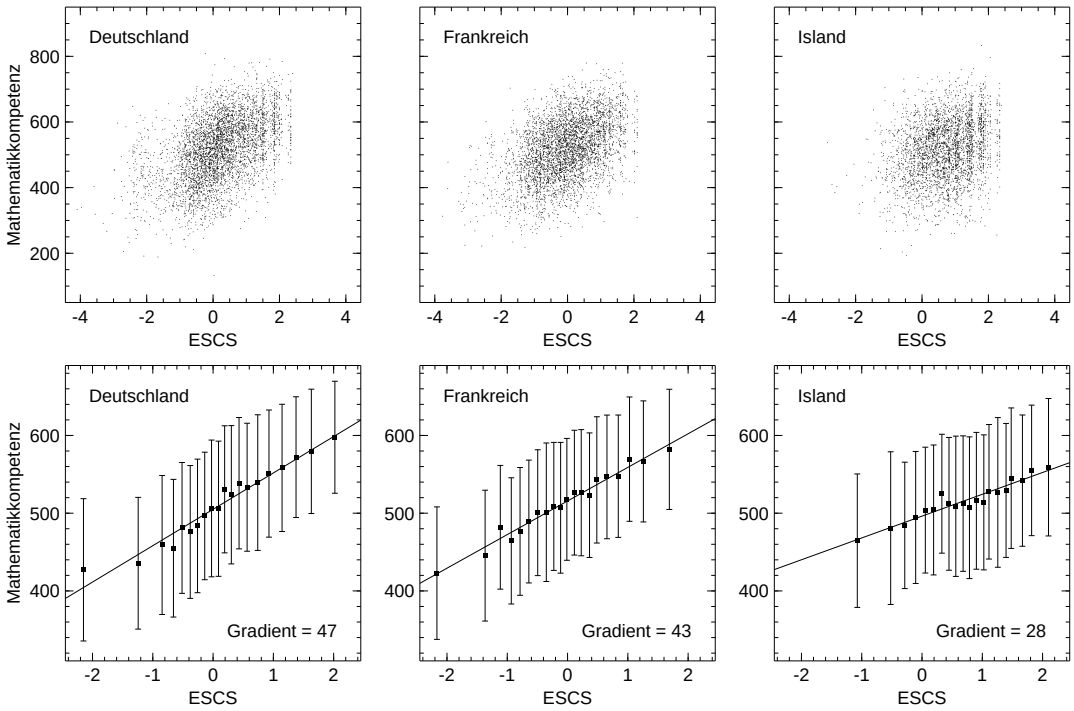

Abbildung 14: Mathematikkompetenz versus ESCS-Index für drei Staaten in zwei verschiedenen Auftragungen. Oben: jeder Punkt repräsentiert einen Probanden. Unten (gestreckte Skalen!): jeder Punkt repräsentiert fünf Prozent der Grundgesamtheit (unter Berücksichtigung unterschiedlicher statistischer Gewichte der Probanden); der vertikale Balken überstreicht nach oben und nach unten je eine Standardabweichung.

> alle Staaten einheitlich quantifiziert. So ist ein internationaler Vergleich der Ergebnisse zwischen den Staaten und über die Zeit hinweg möglich [D03b, S. 248].

Eine weitere Erläuterung

> Würde die soziale Herkunft für den Erwerb von Kompetenzen keine Rolle spielen, wäre es nicht möglich, eine solche Regressionsgerade zu schätzen [D03b, S. 250]

ist blanker Unsinn, denn selbstverständlich funktioniert lineare Regression auch dann noch, wenn als Steigung Null herauskommt.

Im Lesetest von PISA 2000 war für Deutschland der größte soziale Gradient aller OECD-Staaten gefunden worden (D00, S. 386 ff.). Im Mathematiktest von PISA 2003 wurde dieses Ergebnis nicht einmal annähernd reproduziert; Deutschland liegt mit einem Gradienten von 47 auf dem 6. Platz des OECD-Rankings. Die Spanne erstreckt sich von 55 in Belgien und Ungarn bis 28 in Island; mehr als die Hälfte der 29 Staaten liegt in dem engen Bereich zwischen 47 und 41 (LTW, S. 397; D03b, S. 249). Den Berichten zufolge unterscheidet sich

der deutsche Gradient nicht signifikant vom OECD-Mittelwert von 45; allerdings ist dieser Mittelwert falsch berechnet und beträgt tatsächlich 42.[58]

Um zu veranschaulichen, was ein sozio-kultureller Gradient von 47 bedeutet, soll im folgenden Deutschland mit zwei Staaten verglichen werden, die bei ähnlichen Mittelwerten niedrigere Gradienten erzielt haben: Frankreich (43) und, als Extrembeispiel, Island (28). Die obere Zeile in Abbildung 14 zeigt für diese drei Länder einen „plausiblen" Mathematikkompetenzwert pro Proband, aufgetragen gegen seinen ESCS-Wert. In jedem ESCS-Bereich wird eine weite Spanne unterschiedlicher Kompetenzwerte beobachtet. Zu ähnlichen Daten aus PISA 2000 erklären Baumert *et al.*:

> Trotz des systematischen Zusammenhangs ist die Kopplung zwischen Sozialschicht und Kompetenzerwerb begrenzt. Es gibt genügend [!] Jugendliche aus unteren Sozialschichten mit exzellenten Leseleistungen und umgekehrt [D00, S. 387].

Um trotz dieser starken Streuung den systematischen Zusammenhang herauszuarbeiten, sind in der unteren Zeile in Abb. 14 die Schüler in 5 %-Gruppen eingeteilt. Auf den gestreckten Skalen ist der Trend zu zunehmender Mathematikkompetenz mit zunehmendem ESCS-Wert nun sogar für Island zu erkennen. Außerdem ist eine Ausgleichsgerade eingezeichnet, deren Anstieg den sozialen Gradienten darstellt.

Zur weiteren Analyse lohnt es sich, von den Kompetenzwerten auf den Anteil korrekt gelöster Aufgaben zurückzugehen. Wenn man die Auswertung aus der unteren Zeile von Abb. 14 mit Lösungshäufigkeiten anstelle von Kompetenzwerten und einer logistischen Funktion mit variabler Trennschärfe anstelle einer Ausgleichsgeraden wiederholt, dann findet man für Deutschland, Frankreich, Island an der steilsten Stelle Gradienten von 13, 11 und 7 Lösungshäufigkeits-Prozentpunkten pro ESCS-Punkt. Diese Unterschiede zwischen verschiedenen Staaten können nun mit Unterschieden zwischen verschiedenen Testaufgaben verglichen werden.

Abbildung 15 zeigt, mit welcher Häufigkeit deutsche Schüler unterschiedlicher sozialer Herkunft zwei bestimmte Mathematikaufgaben gelöst haben. Auch hier kann der Zusammenhang durch eine logistische Funktion mit variabler Trennschärfe genähert werden. Der Gradient an der steilsten Stelle beträgt 9,4 % pro ESCS-Punkt für „Cube Painting Q1", aber 27,1 % für „Height Q1". Die Varianz der sozialen Selektivität *einzelner* Aufgaben innerhalb eines *einzelnen* Staats ist also erheblich größer als die Varianz der über alle Aufgaben

[58] 42 ist der Mittelwert der Gradienten der 29 Einzelstaaten. 45 ist der Gradient des OECD-weiten Datensatzes (wobei die Schülergewichte so modifiziert sind, dass alle Einzelstaaten gleich stark berücksichtigt werden). Dieser Wert ist sinnlos, weil er international skalierte Kompetenzwerte in Beziehung zu national unterschiedlich normierten und zentrierten ESCS-Indexwerten setzt.

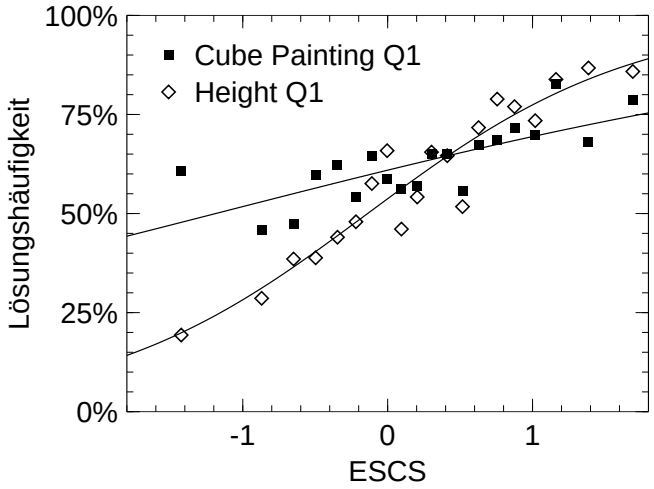

Abbildung 15: Lösungshäufigkeit zweier Mathematikaufgaben in Deutschland als Funktion des ESCS-Indexes. Die durchgezogenen Linien sind Anpassungen mit der logistischen Funktion mit variabler Trennschärfe.

gemittelten Gradienten zwischen *verschiedener* Staaten – erst recht, wenn man Sonderfälle wie Island ausnimmt. Demnach hängen Aussagen über den Zusammenhang zwischen „Bildungsstand“ und „sozialer Herkunft“ nicht nur von den Zufälligkeiten der ESCS-Bewertung, sondern ganz entscheidend auch von der Aufgabenauswahl ab. Abbildung 16 zeigt im direkten Vergleich wie einige Aufgaben im Deutschland, andere in Frankreich die größere soziale Trennschärfe besitzen.

Es ist nicht schwer, Gründe zu finden, warum die Lösungshäufigkeit je nach Testitem unterschiedlich stark mit der sozialen Herkunft der Probanden variiert. Um nur ein Beispiel zu nennen: Die Aufgabe „Daylight“ (Abb. 18) enthält neben anderen Schwächen (Bender 2006) auch einen Übersetzungsfehler: Schulbüchern nach zu urteilen, hätte „hemisphere“ nicht als „Hemisphäre“, sondern als „Erdhälfte“ wiedergegeben werden müssen.[59] Wenn in einer Testaufgabe Vokabular verwendet wird, das im einschlägigen Schulunterricht nicht eingeführt wird, dann hängt der Schwierigkeitsgrad der Aufgabe eben in erhöhtem Maße vom außerschulischen Bildungshintergrund des Schülers ab (vgl. Meyerhöfer 2005, S. 195–198).

5.3 Kompetenzen von Jungen und Mädchen

Als Freudenthal in seiner Analyse der ersten großen Schulsystemvergleichsstudien zu dem Ergebnis kam, dass das Geschlecht der Teilnehmer die einzige

[59] Die Bedeutung der „im jeweiligen nationalen Unterricht akzeptierten fachlichen Sprachkonventionen“ für die „kulturübergreifende Testfairness“ erkennen Baumert *et al.* (2000) durchaus an.

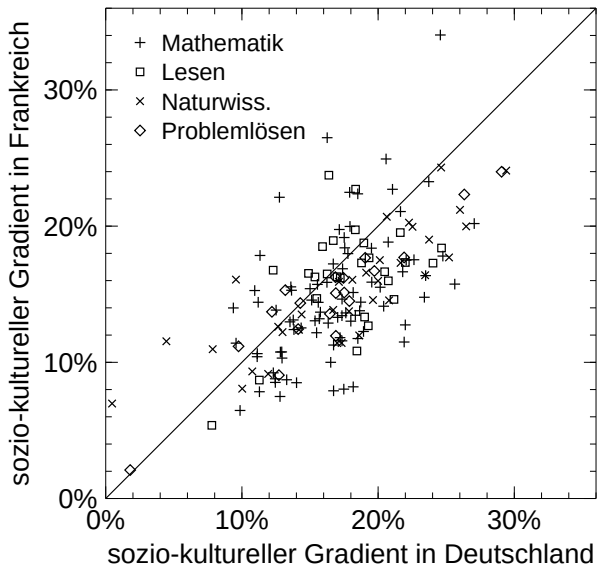

Abbildung 16: Sozio-kulturelle Trennschärfe aller 164 auswertbaren Testitems in Frankreich und in Deutschland. Die 44 Aufgaben oberhalb der Diagonale trennen in Frankreich stärker als in Deutschland; für die übrigen 122 ist es umgekehrt.

wohldefinierte Variable war (1975, S. 177), ahnte er nicht, dass die Genderforschung die Eindeutigkeit auch dieser Variablen in Frage stellen würde. Nachdem jedoch der Versuch, aus den PISA-Daten eine als *gender* deutbare latente Variable zu konstruieren, eine ganze Doktorarbeit lang erfolglos geblieben ist (Burba 2006), erscheint es legitim, bis auf weiteres nur das in herkömmlicher Weise definierte Geschlecht zu berücksichtigen, wie es der Selbstauskunft der Teilnehmer entnommen oder von Amts wegen in den Datensatz eingetragen wurde.

Die offizielle Auswertung stützt sich im wesentlichen auf die nach Staaten und Testgebieten aufgeschlüsselten Differenzen der Kompetenzmittelwerte von Jungen und Mädchen. Tabelle 6 zeigt einen Auszug aus diesem Datensatz. Im Lesetest erzielen die Mädchen in sämtlichen Staaten ein deutlich besseres Ergebnis als die Jungen; die Differenzen liegen zwischen 21 (Mexiko, Niederlande, Südkorea) und 58 (Island) Kompetenzpunkten.[60] In den übrigen drei Testgebieten sind die Differenzen deutlich geringer; überwiegend betragen sie weniger als 10 Punkte. Solche Differenzen sind nicht größer als typische in diesem Aufsatz benannte systematische Unsicherheiten.

Die vorletzte Zeile in der Tabelle gibt einen Anhaltspunkt, wie allein die Unsicherheit der Stichprobenziehung auf Geschlechterdifferenzen durchschlagen

[60]Im OECD-Staatenmittel beträgt der Unterschied 34 Kompetenzpunkte. Der über alle Leseaufgaben gemittelte Unterschied in der Lösungshäufigkeit beträgt 5,3 Prozentpunkte (Tab. 7). Das ist weniger, als nach der in 3.14 angegeben Umrechnung zu erwarten wäre, weil diese nur an der steilsten Stelle des Lösungsprofils gilt.

Tabelle 6: Differenz der Kompetenzmittelwerte zwischen Jungen und Mädchen (positives Vorzeichen: Vorsprung der Jungen). Die beiden Zeilen, die die deutschen Daten nach Lang- und Kurzheften aufschlüsseln, sind neu berechnet; die übrigen Daten finden sich übereinstimmend in den deutschen Berichten (D03a, S. 21; D03b, S. 213).

	Kompetenzpunktedifferenz Jungen − Mädchen			
	Lesen	Mathematik	Naturwiss.	Problemlös.
OECD	–34	11	6	–2
Deutschland	–42	9	6	–6
– nur Langhefte	–36	15	12	–2
– nur Kurzhefte	–21	20	22	2
Österreich	–47	8	–3	–3
Schweden	–37	7	5	–10

kann: Wenn man die in Sonderschulen eingesetzten Kurzhefte ausschlösse (wiederum: ohne zu behaupten, dass das „korrekter“ als die offizielle Auswertung wäre), verschöben sich die Differenzen in Deutschland um 4 bis 6 Punkte zugunsten der Jungen – und das, obwohl *innerhalb* der Sonderschulen die Jungen tendentiell leistungsstärker sind als die Mädchen (letzte Tabellenzeile). Der gegenläufige Nettoeffekt erklärt sich aus der Änderung der Stichprobenzusammensetzung: unter den mit Kurzheft Getesteten waren in Deutschland 71,6 % Jungen. Allein schon die international uneinheitliche, unkontrollierbare Abgrenzung der Stichprobe am unteren Rand des Leistungsspektrums trägt also in der Größenordnung von mindestens fünf Punkten zur systematischen Unsicherheit aller Aussagen über Kompetenzunterschiede zwischen den Geschlechtern bei.

In der offiziellen Auswertung aber werden nationale Mittelwertdifferenzen ab 6 Punkten für signifikant erklärt. Dennoch lassen sich aus dem Zahlenmaterial keine klaren Trends zugunsten bestimmter Ländergruppen ablesen, weshalb die offizielle Interpretation äußerst dürftig bleibt (D03a, S. 20 f.; D03b, S. 212 ff.). Die Ergebnisse im Problemlösetest werden als „Indikator für das kognitive Potential im Bereich Mathematik“ gedeutet, um dann anhand von Differenzen von Differenzen zu argumentieren, dass einige Staaten dieses Potential besser ausschöpfen als andere (D03b, S. 214 f.). Zu den Bereichen Mathematik und Naturwissenschaften heißt es:

> Allerdings zeigt der internationale Vergleich, dass die Abstände sehr unterschiedlich oder gar gegenläufig ausfallen können: Interessante Vergleichsstaaten sind hier Finnland, die Niederlande, Australien, Schweden oder Island. Sie zeigen, dass Schülerinnen und Schüler in diesen Bereichen durchaus ein vergleichbares Kompetenzniveau erreichen können [D03a, S. 20].

Diese Interpretation ist mehr als überraschend, denn die in Tab. 6 wiedergegebenen Daten deuten keineswegs darauf hin, dass Schweden für Deutschland

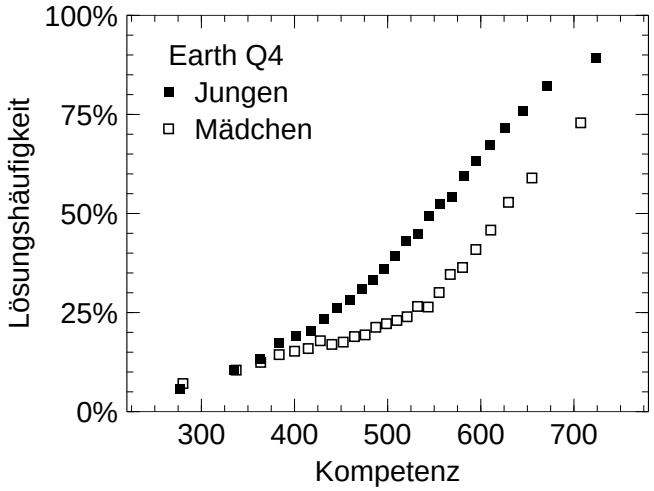

Abbildung 17: Das Lösungsprofil dieser Naturwissenschaftsaufgabe zeigt besonders starke Unterschiede zwischen den Geschlechtern. Im OECD-Staatenmittel wird diese Aufgabe von 42,2 % aller Jungen, aber nur von 27,3 % aller Mädchen richtig gelöst.

ein Vorbild bezüglich der Einebnung von Geschlechterdifferenzen in der Mathematikleistung sein könnte. Transportieren die Auswerter hier ohne jede empirische Basis ihre private Vorliebe für skandinavische Schulsysteme, oder haben sie die Logik der Nullhypothesentests missverstanden und meinen, dass Differenzen signifikanter Differenzen per se signifikant seien?

Eine Koautorin dieser Auswertung hat in ihrer schon erwähnten Dissertation (Burba 2006; Betreuer J. Rost) den Geschlechterunterschied im Naturwissenschaftstest näher untersucht. Das Hauptergebnis ihrer Arbeit,

> dass die Geschlechterunterschiede bei PISA, insbesondere im nationalen Naturwissenschaftstest[,] von untergeordneter Bedeutung sind [S. 152],

beruht auf einem Fehlschluss, denn es stützt sich allein darauf, dass eine bestimmte, blind angewandte Software eine Klasseneinteilung geliefert hat, die nicht in der erhofften Weise mit dem biologischen Geschlecht korreliert, sondern eine andere Verletzung der Eindimensionalität des Rasch-Modells anzeigt.

Es spricht im Gegenteil einiges dafür, dass der Geschlechterunterschied im Naturwissenschaftstest besonders ausgeprägt ist. Im nationalen Test PISA–E 2000 wurden in Deutschland erhebliche Geschlechterdifferenzen je nach naturwissenschaftlichem Fachgebiet gefunden (D00, S. 255). Im internationalen Datensatz von PISA 2003 gibt es eine ganze Reihe von Aufgaben, deren geschlechtsabhängige Schwierigkeit oder/und Trennschärfe das Rasch-Modell verletzt. Die extremsten Fälle sind Naturwissenschaftsaufgaben. Abbildung 17 zeigt ein Beispiel. Dem Rasch-Modell zufolge müssten die Lösungsprofile von Jungen und Mädchen zusammenfallen; unterschiedliche Naturwissenschaftskompetenz dürfte sich nur darin ausdrücken, dass die Perzentilsymbole *entlang* der Rasch-Kurve gegeneinander verschoben sind. Das ist nicht entfernt der Fall; auch ist

der Profilverlauf, besonders deutlich im Fall der Mädchen, nicht mit der Rasch-Antwortfunktion vereinbar.

In Tabelle 7 sind verschiedene statistische Kennwerte zur Geschlechterdifferenz der Lösungshäufigkeiten angegeben und nach den vier Testgebieten aufgeschlüsselt. Die mittlere Differenz $\overline{\delta}$ ist im Lesetest mit 5,3 % bei weitem am größten (das negative Vorzeichen in der Tabelle bedeutet einen Rückstand der Jungen). Ein solcher Leistungsunterschied führt dazu, dass den beiden Subpopulationen unterschiedliche Kompetenzwerte zugeschrieben werden, was mit einem eindimensionalen Item-Response-Modell noch kompatibel ist. Die Verletzung des Modells durch eine Mischung von Aufgaben, die mal die eine, mal die andere Subpopulation begünstigen, äußert sich erst in der *Streuung* der δ. Diese Streuung ist im Naturwissenschaftstest größer ist als in den drei anderen Gebieten. Auch einige weitere in der Tabelle aufgeführte Kriterien bestätigen, dass der Geschlechterunterschied im Naturwissenschaftstest jedenfalls stärker als in Mathematik oder Problemlösen ist.[61]

Die geringe *mittlere* Kompetenzpunktedifferenz zwischen Jungen und Mädchen kommt allein durch Mittelung über eine zufällig recht ausgeglichene Aufgabenauswahl zustande; man könnte nach Belieben andere Naturwissenschaftstests zusammenstellen, die leichter vom einen oder vom anderen Geschlecht

[61] Ponocny (in Neuwirth *et al.* 2006, S. 71 ff.) findet ebenfalls eine „hochsignifikante" Verletzung der Rasch-Modellannahmen im Naturwissenschaftstest, und setzt dann die Unterschiede zwischen österreichischen Mädchen und Burschen in Beziehung zu Unterschieden zwischen verbalem und nonverbalem Antwortformat. Das ist wenig überzeugend, solange nicht ausgeschlossen werden kann, dass der Geschlechterunterschied nicht eher durch die *Inhalte* der Aufgaben als durch das Antwortformat zustande kommt (bei der geringen Anzahl von Aufgaben ist eine zufällige Korrelation zwischen Inhaltsbereich und Format nicht unwahrscheinlich). Ein Grund mehr, warum eine verlässliche Auswertung allenfalls nach Veröffentlichung sämtlicher Aufgaben möglich ist.

Tabelle 7: Statistische Kennwerte zu den Lösungshäufigkeitsdifferenzen $\delta_i = \rho_i(Jungen) - \rho_i(Mädchen)$, ausgewertet jeweils für die Gesamtheit aller Aufgaben $i \in \mathfrak{I}_g$ eines Testgebiets g, im OECD-Staatenmittel.

	Lesen	Math.	Nat.wi.	Probl.lö.
Minimum $\delta_{\min}$	–14,1 %	–6,2 %	–8,2 %	–6,9 %
Maximum $\delta_{\max}$	4,9 %	12,3 %	14,9 %	7,4 %
Median δ_{med}	–4,8 %	2,2 %	2,2 %	–1,1 %
Mittelwert $\overline{\delta}$	–5,3 %	2,6 %	0,9 %	0,6 %
Mittl. absoluter Abstand $\overline{\lvert\delta\rvert}$	5,7 %	3,5 %	4,1 %	3,6 %
Mittl. quadrat. Abstand $\overline{\delta^2}^{1/2}$	6,7 %	4,4 %	5,2 %	4,1 %
Standardabweichung $\overline{(\delta - \overline{\delta})^2}^{1/2}$	4,1 %	3,5 %	5,1 %	4,1 %

gelöst werden. Eine Auswertung, die über die ganze Breite der Naturwissenschaften mittelt, ist von vorneherein zu inhaltlicher Unergiebigkeit verurteilt (vgl. Mathis 2004, S. 145). Didaktisch verwertbare Erkenntnisse kann man allenfalls dann erzielen, wenn man auf der Ebene der einzelnen Aufgaben ansetzt (vgl. Olsen 2005).

6 Zusammenfassung und Bewertung

6.1 Was misst PISA?

Das „literacy"-Konzept und der utilitaristische Bildungsbegriff der OECD bringen mit sich, dass die Testgebiete von PISA ineinander übergehen: Mathematik-, Naturwissenschafts- und Problemlösungsaufgaben enthalten erhebliche Leseanteile, und Leseaufgaben können auch „diskontinuierliche Texte", nämlich Diagramme und Zahlenkolonnen, zur Grundlage haben (Kirsch *et al.* 2002, insbesondere die Beispielaufgaben „Lake Chad" und „Plan International"). Daher überrascht es nicht, dass die Kompetenzwerte in den vier Testgebieten auf Probanden- und auf Staatenebene hohe Korrelationen aufweisen. Somit kann man, in der Sprache der multivariaten Statistik, das Antwortverhalten der Probanden zu einem erheblichen Teil durch *einen* Generalfaktor „erklären".

Rindermann (2006, 2007b) verwendet diesen g-Faktor in einem Datenvergleich zur weltweiten Verteilung von *Intelligenz*. Damit bricht er ein Tabu der Bildungsforschung: der Begriff Intelligenz wird in den PISA-Berichten strikt vermieden – aus gutem Grund, denn kaum eine Regierung dürfte Geld und Untersuchungsgenehmigungen gewähren, um sich über die Intelligenz ihrer Schulbevölkerung informieren zu lassen. Dementsprechend heftig wehrte sich das deutsche Konsortium (Baumert *et al.* 2007, Prenzel *et al.* 2007).

Aus statistischer Sicht, auf Grundlage des internationalen Datensatzes, lässt sich zur Frage, ob der g-Faktor von PISA mit Intelligenz gleichgesetzt werden darf, nichts sagen. Die Repliken des Konsortiums sind aber in anderer Hinsicht interessant. Beide Autorenkollektive argumentieren auf mehreren Ebenen. Eine Ebene sind die Aufgabenanalysen: Da Rindermann sich nur seines eigenen Verstandes und nicht eines Expertengremiums bedient, *verbiete* sich, so Prenzel *et al.*, eine inhaltliche Auseinandersetzung. Aber selbst wenn seine Aufgabenanalyse methodisch adäquat wäre, dürfe man nicht unberücksichtigt lassen, dass die Testkonstruktion qualitätsgesichert, validitätsgeprüft und theoriegeleitet sei.

Dieser selbstgewissen Behauptung stehen die Ergebnisse aus Teil 4 entgegen: Die Vorprüfung der psychometrischen Qualität der Testaufgaben war in PISA unzureichend. Die breite Verteilung der Trennschärfen (Abb. 6 f.), die Überlagerung verschiedener Lösungswege (Abb. 8), nicht-monotone Lösungsprofile (Abb. 9), häufiges Raten (Abb. 10) und die Abhängigkeit mancher Aufgabenschwierigkeiten vom Geschlecht (Abb. 17) passen nicht zu der einparametrigen Item-Response-Theorie, die angeblich die Testkonstruktion geleitet hat.

Auch den Konsortialen ist klar, dass man Zweifel am Output einer Studie nicht allein dadurch ausräumen kann, dass man die Qualität des Inputs bekräftigt (Matth. 7, 20). Um Rindermann auch auf der Ebene der Antwortstatistik zu entgegnen, zitieren Baumert *et al.* Ergebnisse aus der Doktorarbeit von Brunner (2005). Ihre Argumentation ist hart an der Grenze zur Unredlichkeit, denn sie

ist nicht als bloße Analogie kenntlich gemacht: Brunner arbeitet nicht mit dem internationalen, sondern mit dem deutschen Datensatz und unter Verwendung der deutschen Mathematik-Zusatzaufgaben. Diese Aufgaben sind eher kurz, testen spezifische Fertigkeiten und lassen sich schulischen Stoffgebieten zuordnen (Knoche *et al.* 2002, S. 162), während Rindermann auf die Länge, Komplexität und stoffliche Unbestimmtheit der internationalen Aufgaben abstellt. Inwieweit PISA-international neben einem g-Faktor auch gebietsspezifische Kompetenzen misst, müsste empirisch erst noch untersucht werden.[62] Wenn eine solche Untersuchung kein deutlich positives Ergebnis zeitigt, bleibt es bei dem Schluss, dass man sich zwölf der dreizehn Hefte hätte sparen können (W1, S. 149).

Sicher lässt sich sagen, dass PISA nicht *nur* kognitive Fähigkeiten und fachspezifische Kompetenzen misst. Einen quantitativ bedeutsamen Einfluss haben auch Vertrautheit mit dem Aufgabenformat (4.6), landestypische Gegebenheiten (4.7), Testsprache (4.8) sowie Durchhaltevermögen, Zeitmanagement und Teststrategie (4.9). Einige dieser Faktoren kann man zusammenfassend als *Testfähigkeit* (Millman *et al.* 1965, Meyerhöfer 2005) beschreiben, aber selbst die lassen sich nicht *eindimensional* quantifizieren.[63] Die offizielle, eindimensionale Analyse der Schülerleistungen, in der pro Testgebiet nur auf je *einen*, naiv als Kompetenz gedeuteten Faktor geschlossen wird, ist allein schon wegen dieser Störfaktoren inadäquat.

6.2 Wie genau misst PISA?

Allein schon dank der über hundert Fragen des *Student Questionnaire* sind die PISA-Rohdaten ein wertvolles Korpus, das die Lebensumstände Fünfzehnjähriger in weiten Teilen der Welt erschließt. Nur in wenigen Bereichen der Sozialforschung dürfte es international erhobenes Datenmaterial von vergleichbarer Qualität geben.[64] Wenn man die Qualität der PISA-Stichprobe weiter erhöhen wollte, müsste man den Aufwand noch einmal erheblich steigern. Das Stratifizierungsproblem ist wahrscheinlich nicht anders zu bewältigen, als dass man,

[62] Brunner greift auch methodisch zu kurz. Es genügt nicht, zu zeigen, dass man die Testergebnisse vollständiger erklären kann, wenn man den Test in Teile zerlegt und für jeden Teil eine zusätzliche Variable einführt. Um a posteriori zu zeigen, dass die a priori gegebene Zerlegung korrekt ist, müsste man nachweisen, dass jede einzelne Testaufgabe korrekt eingestuft ist. In einer vorläufigen, explorativen Analyse der internationalen Daten finde ich: einige Leseaufgaben testen eine distinkte Fähigkeit, andere eher nicht.

[63] Beispielsweise belegt Österreich im OECD-Vergleich Spitzenplätze in Durchhaltevermögen und Zeiteinteilung (4.9), ist aber letztplaziert bezüglich der Kenntnis des Multiple-Choice-Formats (4.6).

[64] Das französische Unterrichtsministerium, das im Gegensatz zu Deutschland ohne zwischengeschaltetes Forschungsinstitut mit eigenem Sachverstand an PISA beteiligt ist, urteilt jedoch, dass sich aus den Kontextdaten keine verlässlichen Schlüsse ziehen lassen (Cytermann in DESCO 2003, S. 23).

wie jetzt schon in Island und Luxemburg, komplette Jahrgänge testet. Wenn man so weit nicht gehen will, ist Köller (2006a) beinahe zuzustimmen, „dass man es aktuell nicht besser machen kann“. Die Frage ist aber, ob das Beste gut genug ist: ob die Qualität der PISA-Stichprobe ausreicht, um die quantitativen Ergebnisse zu tragen.

Tief im Konzept der Studie verankert ist die Beschränkung auf Jugendliche, die noch zur Schule gehen. Zusammen mit der Wahl des Altersjahrgangs bewirkt sie, dass die Daten nicht voll entwickelter Länder wenig repräsentativ für die Gesamtbevölkerung sind und schon deshalb der angestrebte Schluss auf den „outcome“ von Bildungssystemen nicht zulässig ist (2.1).

Repräsentativität und Vergleichbarkeit der nationalen Stichproben werden außerdem durch den uneinheitlichen Ausschluss behinderter Schüler (2.4), durch die uneinheitliche Einbeziehung von Berufs- und Sonderschulen (2.3, 2.5) sowie durch Schwänzen und Testverweigerung (2.6 f.) beeinträchtigt. Die Ergebnisse aus den USA wurden trotz eines verfehlten Teilnahmequorums in den Datensatz aufgenommen. Fünf Staaten überschreiten außerdem die 5 %-Grenze für Testausschlüsse. Die Stichprobenziehung beruht auf problematischen Ausgangsdaten; sie ist im Detail nicht kontrollierbar und auf verschiedenen Ebenen manipulierbar (2.2). Selbst massive Fehler, wie in Österreich und vermutlich Südtirol, können lange unentdeckt bleiben und werden nur bei entsprechender politischer Interessenlage aufgeklärt. In Südkorea hat die Stichprobenziehung zu einer unglaubwürdigen Ungleichverteilung der Geschlechter und Geburtsmonate geführt (2.8). Der polnische Datensatz zeigt eine Anomalie, die einen Manipulationsverdacht begründet (2.9). Bei den kognitiven Testdaten ist außerdem zu bedenken, dass über die genauen Testabläufe in den Schulen (2.9) und über die Motivation der Probanden (6.1) nur wenig bekannt ist.

Nur wenige dieser Verzerrungen lassen sich näherungsweise quantifizieren. Die in Teil 2 gewagten Abschätzungen sind kompatibel mit folgendem Überschlag: Wenn die schwächsten 5 % von einer (500,100)-Normalverteilung ausgeschlossen werden, hebt das den Mittelwert um fast 11 Punkte, bei Annahme einer realistischeren Abschneidefunktion um größenordnungsmäßig die Hälfte davon.[65] Uneinheitlichkeiten in der Definition der Grundgesamtheit, bei der Stichprobenziehung und bei der Testkomplianz können eine Kumulierung von Verzerrungen bewirken. Eine vorsichtige, für PISA noch günstige Schätzung könnte lauten, dass etliche nationale Mittelwerte um fünf bis zehn Punkte verzerrt sind; wenn verschiedene Ursachen in dieselbe Richtung wirken, kann die

[65] Das Ausschließen der schwächsten 5 % kann man als Multiplikation der Normalverteilung mit einer Stufenfunktion modellieren. Wenn man diese Stufenfunktion durch eine Ogive ersetzt, deren hauptsächlicher Anstieg sich über einen Bereich von etwa 100 Punkten erstreckt, und deren Zentrum so gewählt wird, dass nach wie vor 5 % der Probanden ausgeschlossen werden, dann findet man numerisch oder aus Tabellenwerten der Fehlerfunktion den genannten Anstieg des Mittelwerts um gute 5 Punkte.

Verzerrung in Einzelfällen auch deutlich größer sein. Die Unsicherheit dieser Abschätzung ist von ähnlicher Größenordnung wie die Verzerrung selbst, weshalb eine Korrektur der Daten durch Außenstehende kaum möglich ist.[66]

Diese Abschätzungen stehen nicht im Widerspruch zur Annahme, bei der Stichprobenziehung sei im großen und ganzen sorgfältig gearbeitet worden. Es drängt sich nicht *ein* bestimmter Fehler auf, der die Qualität des gesamten Datensatzes maßgeblich begrenzt, und den nicht gezielt bekämpft zu haben man den Verantwortlichen vorwerfen müsste. Vielmehr scheint die Repräsentativität der Stichprobe durch eine ganze Reihe verschiedener, aber quantitativ ähnlich bedeutsamer Probleme begrenzt. Nur einzelne dieser Probleme zu lösen, würde die Gesamtgenauigkeit nicht nennenswert verbessern. Insoweit ist Köller nochmals ausdrücklich zuzustimmen, dass sich PISA nahe am derzeit Machbaren bewegt.

Wenn man sich darauf einlässt, dass PISA nicht nur die Fähigkeit, PISA-Aufgaben zu lösen, sondern wohldefinierte Kompetenzen messen soll, dann muss man alle mitgemessenen Störfaktoren (Testfähigkeit, Sprache und Übersetzung, curriculare Voraussetzungen) als eine weitere Quelle systematischer Unsicherheit ansehen. Solche Faktoren können Verzerrungen von zig PISA-Punkten verursachen – schon allein die unterschiedliche Ermüdung im Testverlauf kann ja zu Differenzen von über zehn Prozentpunkten in der Lösungshäufigkeit führen.

Somit sind die *systematischen* Unsicherheiten deutlich größer als die in den offiziellen Berichten angegebenen, rein *stochastischen* Standardfehler, die typischerweise zwischen 2 und 3,5 betragen (LTW, S. 92). Daraus ergeben sich die folgenden Schlussfolgerungen:

(1) Die Fehlerangaben in den offiziellen Berichten sind irreführend.

(2) Wenn man mit aller gebotenen Vorsicht (Grabe 2000, Schmidt 2003) eine Gauß'sche Fehlerfortpflanzung, also eine quadratische Addition der systematischen und stochastischen Unsicherheiten, annimmt (ISO 1995), dann scheinen letztere gegenüber ersteren weithin vernachlässigbar. Das aber heißt, und zwar unabhängig von aller anderen Kritik an PISA: der zur Senkung der stochastischen Unsicherheit betriebene Aufwand ist unverhältnismäßig. Eine Halbierung der Stichproben würde die Gesamtunsicherheit nur marginal erhöhen. Daher bleibt es bei dem Schluss, dass man sich Tausende Probanden pro Staat sparen könnte (W1, S. 149). In Anbetracht des derzeit qualitativ Machbaren ist der logistische Aufwand von PISA überzogen.

(3) Wie in 1.2 referiert, ist in den OECD-Berichten genau angegeben, welche Leistungsunterschiede zwischen Staaten statistisch signifikant sind. Berücksichtigt man systematische Unsicherheiten, sinkt der Anteil signifikanter Vergleichspaare drastisch. Es bleibt dabei, dass Finnland signifikant vor

[66] Rindermann (2007b) versucht es immerhin und findet, dass sein Staatenvergleich dadurch an Konsistenz gewinnt.

Deutschland und Deutschland signifikant vor der Türkei und Mexiko liegt, aber bezüglich der meisten anderen Ländern lässt sich über die Position Deutschlands nichts Sicheres sagen (ähnlich Romainville 2002 über die Einstufung Belgiens: „beaucoup de bruit pour rien“).

(4) Wie die systematischen Unsicherheiten auf die übrigen Tertiärdaten durchschlagen, ist von Fall zu Fall einzeln zu untersuchen. Verzerrungen bei der Stichprobenziehung wirken sich zum Beispiel verstärkt auf Varianzen, aber nur schwach auf Korrelationen aus.

An dieser Stelle liegt der Einwand nahe, dass diese weitreichenden Schlussfolgerungen auf einer letztlich spekulativen Abschätzung der systematischen Unsicherheiten beruhen. Dem ist zu entgegnen, dass Aussagen über Messunsicherheiten grundsätzlich, auch in Naturwissenschaft und Technik, Qualitätsurteile sind, die sich zwar auf Kenntnisse objektiver Gegebenheiten stützen, aber „auch immer subjektive Komponenten“ enthalten (Kessel 2001, S. 5):

> Da die Messunsicherheit ein Qualitätsurteil über den Messwert darstellt, bleibt sie einer gewissen subjektiven Einschätzung unterworfen. Sie ist damit nicht unmittelbar messbar. Die Möglichkeit, ihr Beweiskraft zu verleihen, besteht darin, ihren Wert aus den Kenntnissen über den Messprozess schlüssig herzuleiten, und die Kenntnisse zusammen mit dem Weg, wie daraus das Messergebnis gewonnen wurde, transparent zu machen. Nur so wird das gewünschte Vertrauen in die Richtigkeit der Zahlenwerte erzeugt [S. 6]. Dann kann jeder, der es will, sich über die Hintergründe informieren und die Schlussfolgerungen nachvollziehen [S. 5].

Obige Abschätzungen sind transparent; mehr kann Kritik nicht leisten. Genauere Abschätzungen müssen an der Intransparenz der Stichprobenziehung scheitern. Ich bezweifle, dass sich diese Intransparenz durch einen Qualitätssprung der jetzt schon sehr umfangreichen, von den Herausgebern nicht mehr beherrschten Dokumentation heilen lässt; ich vermute auch hier mit Köller, dass man es kaum besser machen kann.

Eine konstruktive Empfehlung muss eher lauten, bei künftigen Untersuchungen den Stichprobenumfang anhand fundierter Kosten-Nutzen-Analysen zu begrenzen, und bei der Veröffentlichung von Ergebnissen konsequent auf systematische Unsicherheiten hinzuweisen — auch auf die Gefahr hin, dass dann nicht mehr viele Ergebnisse übrig bleiben.

6.3 Verzerrungen mit bestimmter Richtung

Bei manchen systematischen Verzerrungen kann man trotz erheblicher Unsicherheit über ihren Betrag doch mit ziemlicher Bestimmtheit angeben, in welche Richtung sie wirken. So lässt sich sagen, dass PISA (1) englischsprachige Staaten überschätzt und (2) die Fähigkeiten der Testteilnehmer unterschätzt.

In der Abbildung wird gezeigt, wie Lichtstrahlen von der Sonne auf die Erde scheinen.

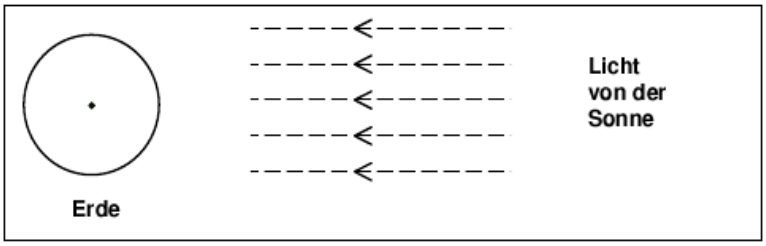

Abbildung: Lichtstrahlen von der Sonne

Nimm an, es wäre der kürzeste Tag in Melbourne.

Zeichne die Erdachse, die nördliche Hemisphäre, die südliche Hemisphäre und den Äquator in die Abbildung ein. Beschrifte alle Teile deiner Antwort.

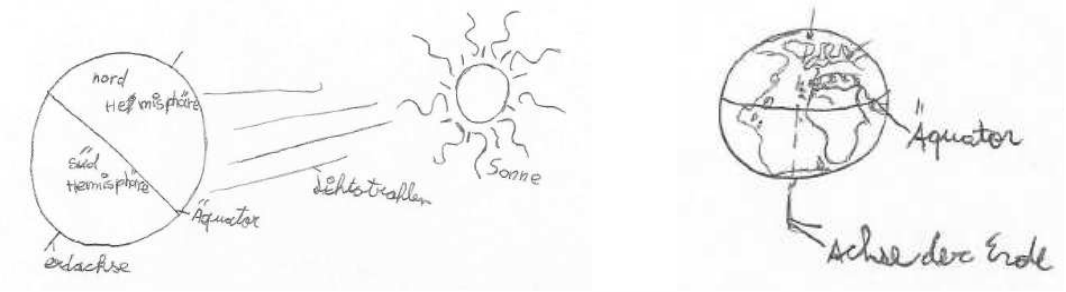

Abbildung 18: Die Naturwissenschaftsaufgabe „Daylight Q2“ und zwei Schülerlösungen aus Luxemburg (Blanke et al. 2004 S. 68, 72). Die Schüler haben die Anweisung „Zeichnen Sie ... ein“ nicht befolgt und stattdessen in den freien Platz unterhalb der Angabe gemalt. Für eine Aufgabe stehen im Mittel 1′40″ zur Verfügung, wovon ein erheblicher Teil zum Lesen langer Einleitungstexte (zur Einheit „Daylight“ über 160 Worte) benötigt wird.

Zu (1): Der Bildungsbegriff der OECD („une optique pratico-pratique très nord-amèricaine“, Romainville 2002), die Herkunft vieler Lesetexte und einer Mehrheit der Testaufgaben aus dem englischen Sprachraum und die Nachbearbeitung sämtlicher Aufgaben durch „professional item writers“ (TR, S. 21) bewirken, dass der kognitive Test Schülern im englischsprachigen Raum vergleichsweise vertraut vorkommt. Auch die Nicht-Anpassung von Personennamen und Ortsangaben wirkt in diese Richtung. In den USA, Kanada und einigen anderen Staaten haben Fünfzehnjährige bereits langjährige Erfahrung mit standardisierten Tests und speziell mit dem Multiple-Choice-Format, so dass von einem erheblichen Vorsprung an *testwiseness* auszugehen ist, der sich empirisch in einer besonders niedrigen Quote von Mehrfachantworten bestätigt (4.6). Die Originalfassung der Aufgaben ist frei von Übersetzungsfehlern, und Lesetexte sind im Englischen kürzer als in manch anderer Sprache (4.8). Zur Überschätzung der Schülerleistungen in den USA und Kanada tragen außerdem

die regelwidrig vielen Ausschlüsse (2.4) sowie möglicherweise die sehr niedrigen Schulteilnahmequoten (2.6) bei.

Zu (2): PISA findet unter erheblichem Zeitdruck statt (4.9). PISA misst nicht Wissen und Können, sondern allenfalls Leistung im Sinne von Arbeit durch Zeit. PISA misst nicht, ob Schüler in der Lage sind, bestimmte Aufgaben zu lösen, sondern wie sie in einer Prüfungssituation mit ganzen Aufgabenbatterien umgehen. Ohne gut geübte Teststrategie (Überspringen langwieriger Aufgaben; Raten) fehlt am Ende Zeit selbst für einfache Aufgaben. Das Nachlassen im Testverlauf (4.9) gibt zumindest einen Anhaltspunkt, wieviel mehr die Schüler unter entspannten Rahmenbedingungen leisten könnten. Wegen der Komplexität vieler Aufgabeneinheiten darf aus einer falschen Lösung nicht auf vollständiges Nichtkönnen geschlossen werden (Aebli 1976, S. 346). Die wenigen zu PISA veröffentlichten Schülerantworten deuten in dieselbe Richtung: in manchen Fällen wurde der Sachverhalt verstanden, aber nicht die Intention der Frage (Blanke *et al.* 2004, z. B. S. 33 f.). Die Korrekturanweisungen sind streng bis pedantisch (Romainville 2002) und in mindestens einem Fall falsch.[67] Weiterhin ist das fehlende Eigeninteresse der Teilnehmer zu berücksichtigen,[68] die in PISA nicht einmal eine Rückmeldung bekommen.[69] Bei PISA kommt die ungünstige Alterstufe hinzu:

> Adolescents, compared with elementary students, were more likely to cheat, to become nervous, to have difficulties concentrating, to guess, and to look for answers that matched the questions without reading the passage. All of these strategies are designed to avoid personal effort and responsibility [...] older students apparently feel greater resentment, anxiety, cynism, and mistrust of standardized achievement tests [Paris *et al.* 1991].

Abbildung 18 zeigt, wie PISA-Teilnehmer ins Malen geraten, statt Leistung im Sinne von Aufgaben pro Zeit zu liefern:

> These students may simply not care whether their scores reflect what they know and can do. Alternatively, because they resent having to take the test, they

[67] „Daylight“ Antwortcode 04: Süden oben einzuzeichnen, ist ein Verstoß gegen eine Konvention, aber kein inhaltlicher Fehler.

[68] In einer Meta-Studie konstatieren Wise und DeMars (2005), dass fehlender Leistungsdruck (*low stakes*) Testergebnisse typischerweise um über eine halbe Standardabweichung drückt (ähnlich auch Bloxom *et al.* 1995). Weitere Verzerrungen ergeben sich durch die Bevorzugung von Aufgaben, die unanstrengend aussehen. Boe *et al.* (2002) quantifizieren das Engagement der TIMSS-Probanden anhand der Quote bearbeiteter Fragen aus dem Questionnaire. Sie deuten diese Quote als „Student Task Persistence“ und finden, dass sie über die Hälfte der Varianz der nationalen Leistungsmittelwerte erklärt.

[69] Wenn standardisierte Tests für Prüfungen (*high stakes*) eingesetzt werden, ist die Geheimhaltung der Aufgaben und damit auch der Korrekturen geradezu ein Rückfall in vormoderne Praktiken (Kohn 2000, S. 18).

> might be noncompliant or intentionally perform poorly [Wise/DeMars 2005, S. 3].

Wer die prozentualen Lösungshäufigkeiten einzelner PISA-Aufgaben wörtlich nimmt, ohne den gesamten Testablauf, das ungewohnte Aufgabenformat, den Zeitdruck und das fehlende Eigeninteresse mitzudenken, unterschätzt die tatsächlichen Fähigkeiten der Schüler und muss erheblichen Bevölkerungsteilen funktionalen Analphabetismus unterstellen.[70]

6.4 Kompetenzstufen

Genau in diesem Sinne wird für Deutschland seit Jahren in unzähligen Variationen verkündet:

> Die katastrophalen Ergebnisse der „Pisa"-Studie betreffen uns alle: Deutschlands Schüler verlernen das Lesen. Die Studie stellt eine faktische Analphabetenrate von fast 22 Prozent fest [Dieter Schormann, Vorsteher des Börsenvereins des deutschen Buchhandels, börsenblatt, 14. 12. 2001].
>
> [...] es kann auf Dauer nicht akzeptiert werden, dass in Deutschland fast jeder vierte Fünfzehnjährige einfache Texte nicht lesen und verstehen kann und günstigenfalls auf Grundschulniveau rechnen kann [Dieter Hundt, Präsident der Bundesvereinigung der Deutschen Arbeitgeberverbände, Deutschlandradio Kultur, 7. 12. 2004].
>
> Eine der dramatischsten Erkenntnisse der Pisa-Studie war doch, dass ein knappes Viertel der 15-Jährigen nicht richtig lesen und rechnen kann – und damit nicht fit für den Arbeitsmarkt ist [Dieter Lenzen, Sprecher eines „Aktionsrats Bildung",[71] Die ZEIT, 8. 3. 2007].

Der Hintergrund ist folgender: Um die Kompetenzpunkteskala über ihre Hilfsfunktion im Staaten-Ranking hinaus „mit Leben zu füllen" (D. Lind *et al.* 2005, S. 84), veröffentlicht das Konsortium die statistischen Ergebnisse von Anfang an zusammen mit einer bestimmten inhaltlichen Interpretation. Diese Interpretation besteht aus einer Einteilung der Aufgabenschwierigkeiten in sieben „Stufen"

[70] Dies dürfte die Tendenz vieler Studien sein. Unplausibel schlechte Ergebnisse der Lesestudie IALS haben zur Aufdeckung einer ganzen Reihe methodischer Mängel geführt (Blum/Guèrin-Pace 2000). Kießwetter (2002) zieht aus seiner Erfahrung in der Begabtenförderung den Schluss, einen Testerfolg als hinreichendes, aber nicht notwendiges Kriterium für eine besondere Begabung anzusehen.

[71] `www.aktionsrat-bildung.de`: „in Bürogemeinschaft mit der Bildungsabteilung der vbw – Vereinigung der Bayerischen Wirtschaft e. V.". Zu dessen Verflechtung mit PISA, Bertelsmann und der ZEIT siehe Bender (2007).

(sechs „Kompetenzstufen“ und eine darunter liegende Inkompetenzstufe, im folgenden als Stufe 0 bezeichnet) und aus verbalen Beschreibungen der jeweiligen Anforderungen.

Wenn die drei Dieter ein knappes Viertel aller Fünfzehnjährigen zu funktionalen Analphabeten erklären, dann beziehen sie sich auf Schüler, deren Kompetenzwerte in die Stufen 0 oder 1 fallen, und übersehen dabei, dass mit „literacy“ schon auf Stufe 1 mehr als nur Lesenkönnen gemeint ist (OECD 2001, S. 47 f.). Die Kieler Projektleitung hat dieser Dramatisierung Vorschub geleistet, indem sie in PISA 2003 den Begriff *Risikogruppe* auf die Stufe 1 ausgeweitet hat (u. a. D03a, S. 6). Das ist international so nicht angelegt, der OECD-Ergebnisbericht spricht von „risk“ nur mit Bezug auf Stufe 0 (LTW, S. 279), und auch im deutschen Bericht zu PISA 2000 war der Begriff noch auf die Stufe 0 beschränkt (D00, S. 120).

Mit Bezug auf diese engere, 10 % aller Fünfzehnjährigen umfassende Risikogruppe wurde dort konstatiert:

> Die von den Lehrkräften vorab als „schwache Leser“ benannten [...] Schüler bilden nur einen kleinen Teil der Risikogruppe. Der größte Teil der [...] Schüler der Risikogruppe wird von den Lehrkräften nicht erkannt [D00, S. 120].

Statt an der Validität ihrer Dateninterpretation zu zweifeln, gefallen sich die Bildungsforscher darin, in zwei Stunden mehr herauszufinden als ein Lehrer in Jahren.

In der öffentlichen Rezeption wird übersehen, dass der relative Anteil von Schülern auf bestimmten Kompetenzstufen primär durch die Konstruktion dieser Stufen bedingt ist (vgl. Kohn S. 14 f.). Die Stufen beruhen auf einer unabhängig vom Testergebnis vorgegebenen Einteilung der Schwierigkeitspunkteskala in sieben Intervalle. Die fünf geschlossenen Intervalle in der Mitte haben alle die gleiche Breite; deren Zahlenwert 62,1 beruht auf substanzloser mathematischer Spielerei.[72] Aufgrund dieser Vorgaben ist zu erwarten, dass der Anteil der Schüler auf den untersten beiden Stufen im OECD-Staatenmittel

$$\int_{-\infty}^{420{,}4} \mathrm{d}\theta^{\mathrm{P}}\, \mathcal{N}(\theta^{\mathrm{P}}; 500, 100) = 21{,}3\,\% \tag{33}$$

beträgt. Tatsächlich liegt das Ergebnis in Mathematik mit 21,5 % leicht darüber, weil Schülerkompetenzen selbst dann nicht exakt normalverteilt sind, wenn eine

[72] Die Erläuterungen (Turner in Adams/Wu 2002, S. 197 ff; TR, S. 253 ff.; LTW, S. 46) sind trotz gegenteiliger Beteuerung („easy-to-understand“) ebenso wirr wie wortreich und verschieben die willkürliche Vorgabe, auf der die Festlegung der Breite beruht, ins Unbestimmte; sie sind nicht einmal widerspruchsfrei: Ein Schüler am unteren Rand einer Kompetenzstufe löst Aufgaben am unteren Rand der Stufe mit der Wahrscheinlichkeit 62 %, Aufgaben in der Mitte mit 50 % (LTW, S. 46). Gemäß (26) müsste die Breite einer Stufe dann $2 \cdot 77{,}89 \cdot \ln(62/38) = 76{,}3$ betragen. Die Angabe, dass Schüler am oberen Rand einer Stufe („masters“) 80 % aller Aufgaben lösen (TR, S. 254), passt auch nicht dazu.

Tabelle 8: Anteil der Probanden in Kompetenzstufe 1 und darunter. Vollständige Angaben sind nur für zwei der vier Testgebiete möglich: für den Naturwissenschaftstest sind die Punktegrenzen nicht dokumentiert; in Problemlösen scheint es nur drei statt sechs Kompetenzstufen zu geben.

Testgebiet	Komp.stufe 1 bis	OECD	Deutschland	Deutschland ohne So.sch.
Lesen	407,5	19,0 %	22,3 %	19,6 %
Mathematik	420,4	21,5 %	21,7 %	19,0 %

Normalverteilung in die Skalierung hineinsteckt wird. Im Lesen ist die Abweichung stärker, weil die Punkteskala an PISA 2000 angeschlossen wurde und sich dabei die Stufengrenzen verschoben haben.

Wie Tabelle 8 auflistet, liegt der Schüleranteil auf den Stufen 0 und 1 in Deutschland in beiden Gebieten, Mathematik und Lesen, bei ungefähr 22 %, also einmal deutlich und einmal knapp über dem internationalen Durchschnitt. Bei Ausschluss der Sonderschüler sänke der Anteil um knappe 3 %, Deutschland läge einmal knapp über und einmal deutlich unter dem OECD-Schnitt. Wie schon in 2.5 betont, ist dies nicht als *Korrektur* des deutschen Ergebnisses zu verstehen, sondern als Beleg für die *Unsicherheit* von Vergleichsaussagen: ob der Anteil besonders schwacher Schüler über oder unter dem internationalen Durchschnitt liegt, hängt empfindlich von unkontrollierbaren Details der Stichprobenziehung ab. Die als solche markierten Sonderschüler gestatten es ausnahmsweise, diese Unsicherheit zu quantifizieren; einige andere in Teil 2 genannte Unschärfen können sich aber ähnlich stark auswirken. Vergleichende Aussagen über den Anteil besonders schwacher Leser dürften zum Beispiel besonders stark dadurch verzerrt sein, dass einige Staaten Legastheniker ausgeschlossen haben. In Anbetracht dieser Unsicherheiten ist keine seriöse Aussage möglich, ob der Anteil von Schülern auf den Kompetenzstufen 0 und 1 in Deutschland bei OECD-weit einheitlicher Testdurchführung leicht über oder leicht unter dem Staatenmittel läge.

Darüber hinaus gibt es zwei fundamentale Einwände gegen die Verbalisierung von Testergebnissen mittels Kompetenzstufen:

(1) Die Klasseneinteilung von Aufgaben und Schülern setzt voraus, dass sich Aufgabenschwierigkeiten und Schülerfähigkeiten auf je einer eindimensionalen Skala anordnen lassen. Im Rahmen der Item-Response-Theorie ist das möglich, solange sich die Testergebnisse mit dem einparametrigen Rasch-Modell beschreiben lassen. Eine wesentliche Stärke von IRT-Modellen besteht darin, die Voraussetzungen ihrer eigenen Anwendbarkeit validieren zu können (Rost 1999). Für PISA ist das Ergebnis negativ (4.2 f.): die

Aufgaben haben sehr unterschiedliche Trennschärfen; das einparametrige Rasch-Modell ist inadäquat. Wenn das Modell dennoch den Daten übergestülpt wird, erhält man Parameterwerte, die sich seriöserweise nicht als Aufgabenschwierigkeit deuten lassen.

(2) Man fragt sich, wie es überhaupt möglich sein soll, die inhaltlichen Anforderungen einer Stufe treffend zu beschreiben und von benachbarten Stufen abzugrenzen, wenn die Lösungshäufigkeiten der Aufgaben in ganz unterschiedlichem Maße von verschiedensten Dimensionen des Testgeschehens abhängen: Kenntnis des Aufgabenformats (4.6), Teststrategie (4.4), Durchhaltevermögen (4.9), Frustrationstoleranz (wenn Probanden nur einen Bruchteil aller Aufgaben zugänglich finden, 3.15), sprachliche Gestalt und kultureller Hintergrund (4.7 f.), unterschiedliche curriculare Voraussetzungen und Lösungswege (4.3, Meyerhöfer 2004b, Bender 2005), sozialer Hintergrund (5.2), Geschlecht (5.3). Die naheliegendste Antwort scheint mir: Die perfekte Strukturiertheit der Kompetenzstufenbeschreibungen deutet auf einen ausgesprochen geringen empirischen Gehalt hin. Die ganze Methodik ist so unempfindlich, dass Unstimmigkeiten in der Zuordnung gar nicht bemerkt werden.

Man unternehme dazu ein Gedankenexperiment: Die Schüler hätten im OECD-Durchschnitt eine halbe Aufgabe pro Mathematikblock mehr gelöst. Wie hätte sich das auf Auswertung und Interpretation der Testergebnisse ausgewirkt? Die Mathematikblöcke umfassen im Durchschnitt zwölf Aufgaben; die Lösungshäufigkeit wäre somit um gute 4 % gestiegen. In PISA-Punkten entspräche das bei festgehaltener Skalierung einem beachtlichen Kompetenzzuwachs um mehr als 16 Punkte; nur noch 17 % der Schüler fänden sich in den Kompetenzstufen 0 und 1. Bei unveränderten Auswerteprozeduren wäre das aber durch die Skalierung kompensiert worden; es wäre letztlich wieder eine Kompetenzverteilung mit Zentrum 500 und Breite 100 herausgekommen; nach wie vor würden rund 22 % der Schüler den Kompetenzstufen 0 und 1 zugeordnet. Hingegen hätte sich die Schwierigkeitseinschätzung der Aufgaben um gut 16 Punkte nach unten verschoben. Ungefähr jedes vierte Aufgabe wäre dadurch in eine niedrigere Kompetenzstufe gekommen; die Anforderungen in den einzelnen Stufen wären entsprechend gestiegen. Die Frage ist nun: Wäre das überhaupt bemerkt worden? Hätten die Interpretationsexperten tatsächlich bestimmte Anforderungen auf einer niedrigeren Stufe angesiedelt, oder Formulierungen geändert? Hätte das deutsche Konsortium bemerkt, dass auf Stufe 1 mehr verlangt wird als schon jetzt, und die Einschätzung „Risikogruppe“ in Frage gestellt?

6.5 Schüler testen Aufgaben

In einem Test werden grundsätzlich nicht nur die Versuchspersonen, sondern immer auch die Aufgaben auf die Probe gestellt. Einem Rohdatensatz sieht man

gar nicht unbedingt an, nach welcher Seite das Untersuchungsinteresse geht; von seiner Struktur her (3.3) sind die Grundmengen $\mathcal{V}$ und $\mathcal{I}$ austauschbar. Das Antwortmodell von Rasch (3.4) überträgt diese Symmetrie auf die Kompetenz- und Schwierigkeitswerte, die bis auf einen Vorzeichenwechsel vertauschbar sind.

Im gängigen Verständnis von PISA dominiert natürlich die Vorstellung, dass Schüler getestet werden. Die Kompetenzstufen beruhen jedoch auf der entgegengesetzen Auswertung: die Verteilung der Schüler auf die Stufen ist im statistischen Mittel immer gleich; die Messung zielt allein auf die Einstufung der Aufgaben. Für *didaktische* Analysen müsste man diesen Ansatz vertiefen und untersuchen, wie Schüler mit einzelnen Aufgaben umgehen (Olsen 2005a). PISA ist allerdings nicht auf eine solche Auswertung hin angelegt:

– Nur ein Bruchteil des anfänglich in Feldtests erprobten Aufgabenmaterials wird im Haupttest eingesetzt (TR, S. 27 f.); der Rest bleibt undokumentiert, obwohl sich aus dem Nicht-Funktionieren bestimmter Aufgaben in bestimmten Ländern vielleicht einiges lernen ließe;
– die Mehrheit aller im Haupttest eingesetzten Aufgaben wird viele Jahre lang geheimgehalten;
– es wird nicht erfasst, wieviel Zeit sich die Probanden für einzelne Aufgaben nehmen;
– die Aufgaben sind so komplex, dass aus einer einzigen Codeziffer, die angibt, ob die Aufgabe richtig, falsch oder gar nicht bearbeitet wurde, herzlich wenig gelernt werden kann.

Die Komplementarität der beiden Sichtweisen, Testen von Schülern und Testen von Aufgaben, kann erhellend wirken; sie hilft, die Tragweite statistischer Aussage einzuschätzen:

– Der Geschlechterunterschied ist im Bereich Lesen „deutlich am größten und am konsistentesten“ (D00, S. 251)? Stimmt auch für PISA 2003, in 26 von 28 Leseaufgaben haben Mädchen die höhere prozentuale Lösungshäufigkeit.
– Die Unterschiede zwischen Jungen und Mädchen in den Naturwissenschaften sind „minimal“ (D03b, S. 212)? Stimmt nicht. Bei einzelnen Naturwissenschaftsaufgaben sind die Unterschiede zwischen den Geschlechtern sogar besonders groß. Nur aufgrund der zufälligen Mischung der Aufgaben mitteln sich diese Unterschiede in der Kompetenzbewertung weitgehend weg (5.3).
– Isländische Schüler bringen signifikant bessere Mathematikleistungen als deutsche? Hängt kritisch von der Aufgabenauswahl ab. Bei 50 Aufgaben schneiden die Isländer besser ab (und zwar in einigen Fällen sehr deutlich, um bis zu 26,2 Prozentpunkte), bei 34 Aufgaben die Deutschen.
– Die Testleistung hängt von der sozialen Herkunft ab? Stimmt, aber die Abhängigkeit ist von Aufgabe zu Aufgabe sehr unterschiedlich ausgeprägt; auch hier hängen Staatenvergleiche in unkontrollierbarer Weise von der Aufgabenauswahl ab.

6.6 Messung von Trends

Die Entscheidung, mit PISA über viele Jahre hinweg Veränderungen messen zu wollen, bringt drei gravierende Nachteile mit sich: Erstens wird der überwiegende Teil der bisher eingesetzten Testaufgaben geheimgehalten, um künftige Durchgänge darauf normieren zu können. Auswertungen auf der Aufgabenebene werden dadurch stark behindert. Wenn man die Qualität der Testaufgaben und ihrer Übersetzungen bewertet, ist man auf die Annahme angewiesen, dass die veröffentlichte Auswahl einigermaßen repräsentativ für den Gesamtbestand ist.

> Aber die Kritik, die häufig geübt wird, funktioniert so, dass man sich eine oder ganz wenige Aufgaben als pars pro toto herausnimmt und daran die ganze Studie misst und schlecht macht. Man hat das Gefühl, dass diejenigen, die diese harte Kritik äußern, nicht genug vom Kuchen abbekommen und dass möglicherweise auch Neid in der Kritik steckt [Köller 2006b].
>
> Mir lagen für Meyerhöfer (2005) alle PISA-Items vor. Die von mir gewählte Methode der kontrastiven Aufgabeninterpretation führt dazu, dass ich die aufgezeigten Probleme für *alle* PISA-Aufgaben behaupte. Qualitativ unterscheiden sich veröffentlichte und nicht veröffentlichte Aufgaben nicht [W. Meyerhöfer, pers. Mitt. Mai 2007].

Zweitens zwingt die Absicht, die Ergebnisse numerisch vergleichbar zu halten, zu weitgehendem Festhalten an der einmal gewählten Testkonzeption (Zwick 1992). Den Verantwortlichen ist es deshalb kaum möglich, Kritik anzuerkennen und umzusetzen (vgl. Bender 2005).

Drittens droht eine Verdoppelung der Kosten: Aus einem Sitzungsbericht der OECD (2005c) geht trotz diplomatischer Sprache deutlich hervor, dass etliche Staaten mit der Wahl der Grundgesamtheit unzufrieden sind und lieber jüngere Schüler testen möchten. Um die begonnene Zeitreihe fortzusetzen, will eine Mehrheit jedoch an der Testung der Fünfzehnjährigen festhalten. So kam der Wunsch auf, PISA in Zukunft für zwei Altersklassen durchzuführen.

Das alles für Datenpunkte, die überwiegend zufallsgesteuert auf und ab fluktuieren. Größere Fluktuationen werden in der offiziellen Auswertung zwar als signifikant bezeichnet; nach allem zuvor Gesagten ist aber klar, dass ungeregelte und undokumentierte Änderungen der Testbedingungen Effekte in ähnlicher Größenordnung bewirken können. Allein schon eine Änderung der *Reihenfolge* von Testaufgaben kann einen steilen Leistungsabfall oder -anstieg von einem Testdurchgang zum nächsten vortäuschen (Zwick 1992). Und selbst wenn sich nach mehreren Dreijahreszyklen ein robuster Trend in den Testleistungen abzeichnete, wäre fraglich, was man daraus schließen dürfte. So heißt es im deutschen Bericht über die einzige „signifikante“ Verbesserung, in Mathematik, die 2003 gegenüber 2000 festgestellt wurde (D03a, S. 9):

> Diese positive Entwicklung könnte auf ein verändertes Problembewusstsein und auf Maßnahmen zurückzuführen sein, die in Deutschland nach TIMSS ergriffen wurden, zum Beispiel durch einen Wandel der Aufgabenkultur [...]

Mit der bisherigen Methodik kann man also nicht einmal klären, inwieweit Verbesserungen der Testergebnisse darauf beruhen, dass *teaching to the test* stattgefunden hat.

6.7 Experten

Die Programmschrift zu PISA 2000 (OECD 1999, S. 10) zählt auf, welche *Indikatoren* die Testung liefern wird, und erklärt dann:

> Although indicators are an adequate means of drawing attention to important issues, they are not usually capable of providing answers to policy questions. OECD/PISA has therefore also developed a policy-oriented analysis plan that will go beyond the reporting of indicators.

Im deutschen Sprachraum ist diese Strategie besonders gut aufgegangen. Die Inszenierung von PISA als ein Nationen-Wettkampf hat enorme öffentliche Aufmerksamkeit erzeugt, den Namen der Studie als eine Marke[73] etabliert und die Macher als Experten akkreditiert, derer man zur Deutung der Ergebnisse dringend bedarf, da man in der Tat aus Punktwerten und Rangplätzen („Indikatoren“) so gut wie nichts lernen kann.

> Statistiken sprechen nicht für sich. Auch die produktiven Impulse aus TIMSS, PISA und IGLU leben nicht von den Zahlen allein, sondern von der Lauterkeit und Klugheit der Leute, die sie interpretieren. Das heißt dann auch: Wenn andere Personen dieselben Zahlen interpretieren, können sie zu ganz anderen Schlussfolgerungen kommen (Ich erinnere an meinen alten Vorschlag, dass die Ergebnisse von Studien aus der Perspektive „stellvertrender LeserInnen“ unterschiedlicher Disziplinen und Positionen vorgestellt werden sollten, um diese Personabhängigkeit der Interpretation auch transparent zu machen [Brügelmann 2006].

[73] Als das IPN im November 2006 eine nationale Studie über Kompetenzzuwächse innerhalb eines Schuljahres herausbrachte, protestierte der *Verband Bildung und Erziehung* gegen „blanke Testeritis ohne ernsthaften wissenschaftlichen Hintergrund“, warnte die KMK davor, sich auf „auf läppischen Nebenschauplätzen zu verirren“ – und forderte, „das Markenzeichen PISA nicht weiter zu beschädigen“ [`http://www.vbe.de/index.php?id=871`]. Für den Vorsitzenden des VBE steht die *Marke* PISA ganz offensichtlich nicht für eine fortlaufende, ergebnisoffene wissenschaftliche Unternehmung, sondern für die „Initialzündung“ von 2001 und die daran geknüpften bildungspolitischen Erwartungen.

Einmal als *Bildungsexperten* anerkannt, beschränken sich die PISA-Macher nicht auf das Erklären von Statistiken. Dabei kommt ihnen ein mediales Bedürfnis nach Wiedererkennbarkeit entgegen, das in allen möglichen Politikbereichen die Figur des universell kompetenten, durch Notorietät legitimierten Großsachverständigen hervorgebracht hat. Beständig auf die Grenzen des fachlich Gesicherten und der eigenen Kompetenz hinzuweisen, ist mit dieser Rolle nicht kompatibel; die Grenze zum Lobbyisten ist unscharf.

Wer sich mit dem PISA-Konsortium anlegt, kommt früher oder später nicht umhin, dessen Verhalten auch auf einer Meta-Ebene zu analysieren (Meyerhöfer 2006a, Rindermann 2007a). Die ständige Berufung der PISA-Macher auf eigene und fremde Expertise muss in eine kritische Untersuchung einbezogen werden, weil sie die interne Konsistenz und die wissenschaftliche Diskutierbarkeit von PISA in Frage stellt:

> The chain of appeals to authority must end somewhere, and, if the whole chain of appeals is to be epistemically sound, it must end with someone who possesses the necessary evidence [Hardwig 2006, S. 329].

Autorität dient in PISA auf verschiedensten Ebenen als Argument. Zum Beispiel, um in Testaufgaben an Grundwissen zu erinnern:

> The President of the Astronomical Society, Mr Perry Vlahos, said the existence of changing seasons in the Northern and Southern Hemispheres was linked to the Earth's 23-degree tilt ["Daylight", LTW, S. 288].

Zum Beispiel in einer unter den Auspizien von Baumert und Köller entstandenen Doktorarbeit, in der Literaturverweise mit dem Hinweis unterlegt sind, der zitierte Autor sei „eine Autorität", „renommiert" oder „namhaft" (Brunner 2005, S. 10, 33, 193). Vor allem aber auf prozeduraler Ebene: Die Berichte (TR, LTW) weisen alle paar Seiten auf die Mitarbeit von Experten hin, darunter „consortium experts", „international experts", „national experts"; „consultants", „individual experts", „expert groups", „expert committees", „expert panels"; „domain-matter experts", „item development experts", „assessment experts", „expert translators", „expert markers"; „trained experts", „suitable experts", „knowledgeable experts", „leading experts"; Experten mit „appropriate expertise", „scientific expertise", „technical expertise" und Experten ohne spezifizierte Expertise – letztere genießen laut ISEI immerhin denselben sozioökonomischen Status wie Astrologen.

Wahrscheinlich sind die Verweise auf die eingebrachte Expertise in jedem Einzelfall als Beleg für hohe Präzision und konsequente Wissenschaftlichkeit gemeint. In ihrer Häufung aber sind sie ein Indiz für ein fundamentales methodisches Problem. Dieses Problem ist in der Verknüpfung quantitativer und qualitativer Verfahren zu verorten und spiegelt sich in der Arbeitsteilung von Psychometrikern und Didaktikern wieder. Nicht zufällig wird in den internationalen Berichten an keiner Stelle auf *psychometrische* Expertise verwiesen:

diese eine Expertise besitzen die Herausgeber selbst. Man beruft sich nur auf Expertise, die man selbst *nicht* hat.

Schon Freudenthal sah die Arbeitsteilung als ein Kernproblem:

> What happens in educational research looks as though in natural science it would have become a habit that – because of the importance of mathematics as a tool – all research is done by mathematicians, who for experiments, if need be, would hire some analysts, laboratory assistants, and stablemen. Fortunately science is not run this way. Otherwise instead of science we would have orgies of bad mathematics [1975, S. 178].

„Bad mathematics“, weil alles Skalieren und Kalibrieren des Outputs nichts nützt, wenn der Input nicht stimmt. Ein Test kann nicht besser sein als seine Aufgaben. In Großstudien wie PISA wird versucht, die Subjektivität der Aufgabenauswahl durch Institutionalisierung des Expertentums (groups, committees, panels) in den Griff zu bekommen. Die bisher veröffentlichten Aufgaben belegen das Scheitern dieses Ansatzes. Zugleich wird Verantwortung verwischt: kein Fachdidaktiker knüpft seine Reputation an einzelne Testaufgaben. Die Experten legitimieren sich allein durch das, was sie *vor* PISA geleistet haben.

Die öffentliche Rezeption von PISA beruht nicht unmittelbar auf den internationalen Berichten, sondern wird durch nationale Auswertungen und Interpretationen vermittelt. Auf dieser Ebene liegt die Deutungshoheit nicht mehr bei Psychometrikern, sondern bei Pädagogen, Psychologen oder Soziologen, die aufgrund ihrer Vermittlungsleistung von den Medien als Bildungsexperten wahrgenommen werden. Die beteiligten Fachdidaktiker treten wesentlich zurückhaltender auf und warnen die Fachöffentlichkeit vor vereinfachenden Interpretationen:

> Eine Beurteilung der Analysen der Tests ist nur möglich, wenn man neben den konzeptionellen Vorstellungen und den zu untersuchenden Fragestellungen bei der Entwicklung des Tests mit den Modellvorstellungen vertraut ist, die messtheoretisch die geplanten Analysen der erhobenen Daten im Blick haben. Verständnisschwierigkeiten bei der Diskussion von „Ergebnissen“ beruhen oft auf fehlender Gesamtschau beider Komponenten [Knoche *et al.* 2002, S. 160].

Das ist zwar sprachlich verunglückt, lässt aber einen Gedanken durchscheinen, dem man, die Anführungszeichen eingeschlossen, nur zustimmen kann: Wer Tertiärdaten deuten will, sollte verstehen, auf was für Primärdaten und auf was für Auswerteschritten sie beruhen.

Der vorliegende Aufsatz dient nicht zuletzt dem Ziel, dieses Verständnis zu fördern. Er schließt Lücken in der offiziellen Dokumentation und zeigt Ungenauigkeiten und Fehler in zahlreichen Publikationen der PISA-Experten auf:

– Adams selbst beschreibt die Skalierung lückenhaft, fehlerhaft, undidaktisch und unsouverän;
– andere Mitarbeiter von ACER fassen das Skalierungsverfahren in einem späteren Kapitel fehlerhaft zusammen;

- die Schweizer Projektleitung meint, es werde ein zweiparametriges Item-Response-Modell verwendet (alles 3.2);
- niemand stört sich daran, dass die Modellierung des Schülerverhaltens offenkundig inadäquat ist (4.2 ff.);
- die deutschen Mathematik-Experten müssen *raten*, was für eine Kompetenzverteilung in der Skalierung vorgegeben wird;
- genau beschreiben sie nur ein *nicht* angewandtes Schätzverfahren, und vergessen dabei die a-priori-Verteilung;
- sie werfen Begründungen für eine Item-Response-Skalierung, für eine Maximum-Marginal-Likelihood-Schätzung und für die Verwendung plausibler Werte durcheinander (alles Anhang E);
- ähnlich wie Köller, der Modellwahl und Plausible-Werte-Methode in unzutreffenden Zusammenhang bringt (Anhang D).

Mangel an fachlicher Souveränität zeigt sich auch in den Reaktionen auf W1:

- Auf die Vermutung, Inkonsistenzen zwischen Datensatz und Dokumentation könnten auf einen Programmfehler zurückzuführen sein, antwortet Köller mit einem naiven Glaubensbekenntnis (Anhang D);
- Prenzel und Walter finden zwar die Ursache der Inkonsistenz; beim Versuch, eine unzutreffende Rekonstruktion des Skalierungsverfahrens als einen Rechenfehler erscheinen zu lassen, verrechnen sie sich dann aber selber (Anhang B);
- ein eindeutiger Fehler, eine mit der Item-Response-Skalierung unvereinbare Abweichung zwischen unterschiedlich gewichteten Lösungsprofilen, ist hingegen unkommentiert, wahrscheinlich also unbemerkt geblieben (Anhang C).

Das ist die Kehrseite der Arbeitsteilung: Viele PISA-Experten verstehen den Datenfluss von der Item-Response-Modellierung bis zur Generierung plausibler Werte allenfalls oberflächlich. Kein Wunder, dass sie tendenziell die Aussagekraft numerischer Ergebnisse überschätzen. Selbst manche Statistiker unter den Experten trauen sich kein eigenes Urteil über die Validität der Datenaufbereitung zu. Das Schweizer Bundesamt für Statistik antwortete auf mein erstes Manuskript:

> Ihre geballte Ladung von Kritikpunkten an PISA erstaunt mich. Leider sind wir zur Zeit sehr beschäftigt, so dass ich Ihr Dokument nicht im Detail lesen kann. [...] Generell möchte ich anmerken, dass Ihre Kritik nur dann ernst genommen werden kann, wenn Sie alle aufgeführten Punkte und Vermutungen von den Hauptverantwortlichen des PISA-Konsortiums und der OECD, das sind Ray Adams und Andreas Schleicher, überprüfen und wenn möglich kommentieren lassen [C. Zahner-Rossier, Mail vom 28. 3. 2006].

Diese Kommentierung ist inzwischen öffentlich erfolgt:

> „Ganz offensichtlich kennt Herr Wuttke das Pisa-Programm nicht wirklich", bügelt Pisa-Koordinator Andreas Schleicher den Tadel des Physikers ab. „100 Ex-

> perten in 30 Ländern arbeiten seit Jahren an diesem System.“ [Spiegel Online 8. 11. 2006].

So erweist sich die Berufung auf Expertise als zirkulär.

Anhänge

A Konkordanz 1./2. Auflage

Für Leser der ersten Auflage dieses Buchs ist in der folgenden Tabelle zusammengestellt, inwieweit ich die dort (W1) erhobenen Kritikpunkte aufrecht erhalte.

W1	Kritik	hier	Status
§ 1	uneinheitliche Einschreibquoten	2.1	präzisiert
§ 2	fehlerhafte Stichprobenziehung	2.2 f.	erheblich erweitert: auch in USA unglaubwürdig; massive Probleme in Österreich; fehlende Schulpflichtige in Südtirol
§ 3	uneinheitliche Ausschlüsse	2.4	unverändert
§ 4	uneinheitliche Einbeziehung von Sonderschulen	2.5	Befund unverändert; Implikationen präziser diskutiert
§ 5	Korrelelation von Teilnahmequote und Testleistung	2.6 f.	präzisiert; weitere Indizien für Korrelelation von Teilnahmeneigung und Leistungsfähigkeit
§ 6	nicht repräsentative Geschlechterverteilung	2.8	Fehler in Korea nach wie vor sehr wahrscheinlich
§ 7	uneinheitlicher Umgang mit unvollständigen Heften	2.9	Anomalie in Kanada erklärt; Manipulationsverdacht gegen Polen unbeantwortet
§ 8	uneinheitlicher Testtermin	—	zurückgezogen: quantitativ wahrscheinlich unbedeutend (Schulz Punkt 6)
§ 10	obskure Dokumentation	3.2	erweitert
§ 11	Lösungshäufigkeiten nicht reproduzierbar	3.13	Befund unverändert und unerklärt
§ 11	Lösungshäufigkeit vs. Schwierigkeit nicht monoton	3.14	Befund unverändert und unerklärt
§ 11	Abweichung zwischen Dokumentation und Daten	B	Abweichung besteht; Erklärungsversuch (Programmierfehler) war falsch
§ 12	Mittelwerte hängen empfindlich von einzelnen Aufgaben ab	3.14	Befund unverändert; vertiefte Diskussion in 6.5
§ 12	Einfluss der Sprache schon wegen Textlänge	4.8	präzisiert und erweitert
§ 13	undokumentierte nachträgliche Umskalierung	3.11	Rekonstruktion war richtig; Umskalierung ist an entlegender Stelle ansatzweise dokumentiert
§ 14	unerwartet starke Korrelation Teilnahmequote $\leftrightarrow$ Schulmittelwert	—	zurückgezogen: Fehler in meiner Datenauswertung (Anh. C)
§ 14	Verstimmung der Skalen	—	zurückgezogen; vergleiche aber viel stärkere Verstimmung durch inadäquate Modellierung (Abb. 11)

W1	Kritik	hier	Status
§ 15	unterschiedl. Trennschärfen nicht modelliert, Aufgaben-Ranking willkürlich	4.2	unverändert
§ 16	unterschiedl. Lösungswege widersprechen Kompetenzstufen	4.3	unverändert
§ 16	einzelne Items völlig missglückt	4.3	unverändert
§ 17	Raten nicht modelliert	4.4	unverändert
§ 17	unterschiedliche Testgewohnheit	4.4	unverändert
§ 18	unterschiedliche Vertrautheit mit Multiple-Choice	4.6	vertieft
§ 18	einzelne MC-Aufgabe völlig missglückt	4.6	unverändert
§ 19	Leseaufgaben messen Weltwissen	4.7	unverändert
§ 20	Einfluss von Sprache und Kultur	4.8	vertieft
§ 21	unterschiedliche Ermüdung	4.9	präzisiert

In der Zusammenfassung hatte ich die Kritik an der Punkteberechnung in zehn Vorwürfe gefasst (W1, S. 145 f.). Errata ergeben sich aus denen zu § 11 und § 14 sowie aus der Klärung der Logik der Item-Response-Kalibrierung (3.5):

Nr.	Kritik	Status
1	Britische Daten in Kalibrierung einbezogen	zeigt prozedurale Probleme, ist aber quantitativ unbedeutend (Fußn. 75)
2	Sonderschulen nicht in Kalibrierung einbezogen	zeigt theoretische Begrenzungen, ist aber quantitativ unbedeutend (Fußn. 75)
3	Probandengewichte nicht in Kalibrierung einbezogen	zurückgezogen (Fußn. 28)
4a	Aufgabenkalibrierung schlecht dokumentiert	noch mehr Mängel aufgedeckt (3.2, Anh. B)
4b	Lösungshäufigk. vs. Schwierigkeit nicht monoton	Befund unverändert und unerklärt (3.14)
5	Bevölkerungsmodell in Aufgabenkalibrierung nicht berücksichtigt	zurückgezogen: Modell wird berücksichtigt, allerdings nicht mit Breite 100 (3.6, Anh. B)
6	unbegründete Gleichsetzung der Rasch-Trennschärfe mit Bevölkerungs-Standardabweichung	zurückgezogen: Umskalierung wirkt auch hier (3.6, insbes. Fußn. 26)
7	Bestimmung der plausiblen Kompetenzwerte nicht dokumentiert	präzisiert: konditionierende Hintergrundvariablen-Kombinationen nicht dokumentiert (3.7)
8	nachträgliche, undokumentierte Umskalierung der Kompetenzskala	bestätigt (Anh. B)
9	Rasch-Modell empirisch inadäquat	unverändert (4.2 ff.)
10	offizielle Aufg.schwierigk. verstoßen gegen 62 %-Verankerung	unverändert (4.2)

Die Fehler aus §§ 11,14 wirken sich auch auf den zweiten Absatz der zusammenfassenden Kritik an den Kompetenzstufen aus (W1, S. 147). Bereits im darauf folgenden Absatz war jedoch schon antizipiert, dass diese technischen Aspekte für die Bewertung nicht ausschlaggebend sind („auch bei einer kompletten Neuskalierung [...] wären die Kompetenzstufen nicht zu retten“).

Zusammengefasst stellt sich der Stand der Debatte so dar:

– Von den Einwänden gegen die Repräsentativität der Stichprobe (Teil I von W1, jetzt Teil 2) ist ein einziger befriedigend beantwortet (fehlende Questionnaires in Kanada, 2.9); weitere Einwände sind dazugekommen (v. a. 2.3).
– Die statistischen Auswertungen einzelner Testaufgaben (Teil III von W1, jetzt Teil 4) sind nicht in Frage gestellt worden. Köller (2006a, Punkt 5) verteidigt die Wahl des einparametrigen Rasch-Modells als eine „Frage der Weltanschauung“; dazu Anhang D. Der Hinweis auf die unterschiedliche Leistungsabnahme im Testverlauf wird von Prenzel und Köller missverstanden; dazu 4.9.
– Die Rekonstruktion des Skalierungsverfahrens in Teil II von W1, unternommen, um die Aufgabenauswertung in Teil III auf eine sichere Grundlage zu stellen, war an einer Schlüsselstelle falsch (Anh. B). Ursache waren einerseits genau benennbare Mängel des Technischen Berichts, andererseits die nahezu perfekte Übereinstimmung eines fälschlich angenommenen Funktionsverlaufs mit dem zutreffenden (Abb. 3). Wegen dieser Übereinstimmung bleiben die übrigen Schlussfolgerungen von W1 unberührt.
– Dass Prenzel und Walter in ihrer Entgegnung zu Teil II einen ebenfalls falschen, auch empirisch unzutreffenden Funktionsverlauf angeben (Abb. 3, Anh. B), ist ein Indiz dafür, dass die Skalierung auch vielen Mitverantwortlichen nicht restlos klar ist. Weitere Indizien dafür sind falsche Darstellungen in offiziellen Berichten (3.2) sowie das Missverstehen der plausible-Werte-Methode durch Köller (Anhang D) und die deutschen Mathematikdidaktik-Experten (Anhang E).

Neben der Berücksichtigung dieser Punkte geht der hier vorliegende Aufsatz vor allem in zwei Teilen weit über W1 hinaus:

– Teil 3 ist weitgehend neu und enthält nun eine selbstkonsistente, ohne Hinzuziehen weiterer Literatur nachvollziehbare Rekonstruktion des in PISA angewandten Skalierungsverfahrens; eine solche Beschreibung hat das Konsortium bis heute nicht geliefert. Wegen Lücken der offiziellen Dokumentation können allerdings weder Lösungshäufigkeiten noch Aufgabenschwierigkeitsparameter unabhängig reproduziert werden. Der Umrechenfaktor aus W1, der verdeutlicht, wie empfindlich quantitative PISA-Ergebnisse von einzelnen Testaufgaben abhängen, wird bestätigt.
– Im neuen Teil 5 werden nun auch Aussagen über die Abhängigkeit der Testleistung vom sozialen Hintergrund und vom Geschlecht untersucht.

Die Einleitung ist leicht, der Schlussteil stark überarbeitet; die Anhänge sind neu. Aus Platzgründen sind die inhaltlich unverändert gültigen Abbildungen 1, 2 und 5 aus W1 durch Text ersetzt worden.

B Erratum zu W1: Umskalierung falsch rekonstruiert

Ausgangspunkt für meine erste Wortmeldung zu PISA waren die nun in den Abschnitten 4.6 bis 4.9 beschriebenen Hinweise auf eine Mehrdimensionalität des Testgeschehens. Beim Redigieren dieser Befunde stieß ich dann auf weitere Ungereimtheiten. Die Auswertungen für die Teile I und III von W1 (nun Teile 2 und 4) erforderten immer wieder Umrechnungen zwischen Lösungshäufigkeiten und Kompetenzwerten. Diese Umrechnungen in vernünftiger erster Näherung durchzuführen, war zwar einfach; sie sauber zu dokumentieren aber unerwartet schwierig, denn vermeintlich äquivalente Ansätze lieferten unterschiedliche Ergebnisse. Die Diskrepanzen, typischerweise im Bereich von 10 %, waren zwar bedeutungslos für Schlussfolgerungen über systematische Verzerrungen durch uneinheitliche Stichproben und uneinheitliches Funktionieren der Testaufgaben, konnten aber bei der Dokumentation meiner Ergebnisse nicht unberücksichtigt bleiben.

Aus der Beschreibung dieser Inkonsistenzen ist Teil II von W1 entstanden (nun durch Teil 3 ersetzt). Bei dessen Ausarbeitung habe ich mich auf die Dokumentation der Skalierung in Kapitel 9 des Technischen Berichts verlassen. Da der Technische Bericht ohne Erläuterungen zwischen einer internen „Logit"-Skala (Mittelwert der Kompetenzverteilung bei 0, Breite circa 1) und der nach außen kommunizierten Punkteskala (Mittelwert 500, Breite 100) hin- und herspringt, bin ich von einer trivialen Umrechnung

$$P = 500 + 100J \tag{34}$$

ausgegangen (W1, Fußn. 9) und habe dementsprechend auch unterstellt, dass das Bevölkerungsmodell, nach außen stets als eine Normalverteilung der Breite 100 beschrieben, intern eine fixe Breite 1 hat.

Aus Inkonsistenzen in den vom Konsortium publizierten Lösungshäufigkeiten, Aufgabenschwierigkeiten und Schülerkompetenzen hatte ich dann erschlossen, dass (1) das Bevölkerungsmodell bei der Bestimmung der Aufgabenschwierigkeiten nicht korrekt berücksichtigt worden sein kann (W1, S. 121), und (2) eine nachträgliche nichttriviale Umskalierung der Schülerkompetenzen stattgefunden haben muss (W1, S. 125). Auf dieser Grundlage habe ich die *Hypothese* (W1, S. 121 f.) geäußert, dass die Diskrepanz (1) möglicherweise auf einen Programmierfehler zurückgeht, der unbemerkt bleiben konnte, weil seine Auswirkungen, mit Ausnahme einer verstimmten Aufgabenschwierigkeitsskala, durch die Umskalierung (2) neutralisiert wurden.

Keinen Versuch unternommen zu haben, diese Hypothese mit ACER abzuklären, war ein bedauerlicher Fehler, an dem Termindruck und eine triviale Kommunikationspanne Anteil hatten. Dadurch ist erhebliche öffentliche Aufmerksamkeit auf eine randständige technische Frage fehlgelenkt worden. Knapp zwei Wochen nach den ersten Presseberichten gelang es Prenzel und Walter (2006), die Inkonsistenzen aufzuklären. Ihr Kurztext sowie weitere, in Teil 3 genannte Quellen erweisen meine Erklärung zu (1) und folglich auch die Annahme eines Programmierfehlers als falsch. Die Unstimmigkeiten zwischen Datensatz und Dokumentation erklären sich vielmehr wie folgt:

Zu (1): Die Breite des Bevölkerungsmodells wird in PISA *nicht* als fix angenommen, sondern simultan mit den Aufgabenschwierigkeiten geschätzt (Parameter δ in 3.5). Eine entgegenlautende Angabe im Technischen Bericht, derzufolge die Aufgabenschwierigkeiten *simultan* mit den *individuellen* Personenparametern geschätzt wurden (TR, S. 250), ist demnach falsch.

Zu (2): Es findet in PISA tatsächlich eine nachträgliche Umskalierung statt. Sie wird, wie oben in Gleichung (24) beschrieben, bei *jeder* Umrechnung zwischen internen und externen Einheiten angewandt; die Annahme einer trivialen Umrechnung (34) war falsch. In der abschließenden Aufzählung von drei Auswerteschritten in Kapitel 9 des Technischen Berichts (TR, S. 122; W1, S. 125) fehlt die Umskalierung. Sie wird erst am Ende des Ergebniskapitels 13 angegeben, ohne jeden Vorausverweis aus dem Skalierungskapitel 9 und ohne jede nähere Erklärung, wo die Koeffizienten herkommen.[74]

Abbildung 3 zeigt, warum die falsche Rekonstruktion aus W1 plausibel scheinen konnte: sie beschreibt den Zusammenhang zwischen offiziellen Lösungshäufigkeiten ρ_i und Aufgabenschwierigkeiten ξ_i so gut, wie das angesichts dessen unerklärter Streuung nur möglich ist, und ist von dem nach korrigierter Rekonstruktion zu erwartenden Zusammenhang (29) nicht zu unterscheiden. Hingegen ist der von Prenzel und Walter angegebene Zusammenhang falsch. Bei dem Versuch, meine unzutreffende Rekonstruktion als einen Rechenfehler darzustellen, haben sie sich nämlich verrechnet: Sie zitieren Gleichung (W1:4) und behaupteten, mir sei „ein Fehler unterlaufen“; ich hätte dort wie auch schon in (W1:1) die Breite 77,89 anstelle der Breite 100 einsetzen müssen. In der Notation von Teil 3 läuft das auf den Zusammenhang

$$\rho(\xi^{\mathrm{P}}) = \int \mathrm{d}\theta^{\mathrm{P}}\, A^{\mathrm{P}}_{\mathrm{Rasch}}(\text{richtig}, \xi^{\mathrm{P}}, \theta^{\mathrm{P}})\, \mathcal{N}(\theta^{\mathrm{P}};\, 500,\, 100) \tag{35}$$

[74] Prenzel und Walter behaupten, ich hätte die Umrechnung aus den Seiten 412/413 des Technischen Berichts erschließen müssen. Sie sagen leider nicht, wie. Die genannten Seiten enthalten eine nackte Tabelle, ohne jeden Rückverweis. Genau aus diesen Tabellendaten hatte ich erschlossen, dass die Skalierung nicht so erfolgt sein kann, wie sie in Kapitel 9 des Technischen Berichts beschrieben wird. Überdies ist, wie in 3.6 beschrieben, schon die elementarste Datenspalte jener Tabelle, „international correct“, nicht nachvollziehbar.

hinaus, der von der korrekten Gleichung (29) durch die falsche Zentrierung der Normalverteilung bei 500 abweicht. Dabei haben Prenzel und Walter übersehen, dass sich die Transformation (24) von der trivialen Umrechnung (34) nicht nur durch einen Stauchungsfaktor 77,89/100, sondern auch durch eine affine Verschiebung um 10,47 unterscheidet, die sich zwar im Antwortmodell, aber nicht im Bevölkerungsmodell $\mathcal{N}(\theta)$ weghebt. Infolgedessen liegt die von ihnen konstruierte Kurve systematisch unter den Daten.

C Erratum zu W1: Falsche Gewichte in Lösungsprofil

Abbildung 6 in W1, hier reproduziert als Teil von Abbildung 19, zeigt die Lösungshäufigkeit als Funktion der plausiblen Kompetenzwerte für die Aufgabe „Growing Up Q3“. Jeder Datenpunkt repräsentiert 4 % der OECD-Stichprobe. Gezeigt sind zwei verschiedene Verläufe: mit und ohne Berücksichtigung der Probandengewichte (geschlossene/offene Symbole). Beide Kurven liegen im mittleren Bereich in vertikaler Richtung um fast 5 % auseinander, was als Hinweis darauf interpretiert wurde, dass die Aufgabenschwierigkeiten unter anderem auch dadurch verzerrt werden, dass bei ihrer Bestimmung die Probandengewichte unberücksichtigt bleiben.

Diese Daten sind fehlerhaft; die vertikale Differenz ist durch unterschiedliche, inkorrekte Staatengewichte zustande gekommen. Bei einheitlicher Durchführung der OECD-Mittelung (hier und auch sonst: Gleichgewichtung aller Staaten) wirkt sich das Probandengewicht wesentlich schwächer aus. Vergleichsweise den stärksten Einfluss hat es bei den untersten Quantilen, und auch dort nicht in vertikaler Richtung, sondern als Verschiebung entlang der Kurve (Abb. 19). Der Grund ist folgender:

Das Probandengewicht nicht zu berücksichtigen, läuft darauf hinaus, den Anteil schwacher Schüler noch weiter zu unterschätzen, als das aus verschiedenen, in Teil 2 erklärten Gründen, ohnehin schon der Fall ist. Dadurch steigt die durchschnittliche Testleistung in den unteren Quantilen. Zugleich steigen aber auch die jeweiligen Kompetenzdurchschnitte. Wenn ein Item-Response-Modell, wie in dieser Aufgabe, das Schülerverhalten im großen und ganzen korrekt beschreibt, dann wirkt sich die Verschiebung der Datenpunkte entlang der Modellkurve nicht auf die Parametrierung der Aufgabenschwierigkeit aus.[75]

[75] In W1, S. 145 hatte ich kritisiert, dass bei der Skalierung der Aufgabenschwierigkeiten Großbritannien berücksichtigt, Sonderschulen aber ausgeschlossen wurden, obwohl Großbritannien von den Ergebnisdarstellungen disqualifiziert war und Sonderschulen in die Berechnung sämtlicher Statistiken einbezogen wurde. Solange das verwendete Item-Response-Modell funktioniert, wirken sich diese Inkonsistenzen nicht verzerrend auf die Schätzung der Aufgabenparameter aus. Da das Schülerverhalten in PISA deutlich vom verwendeten Rasch-Modell abweicht, sind Verzerrungen zwar doch möglich, aber nur als Effekte höherer Ordnung, die quantitativ gegenüber anderen Fehlerquellen wahrscheinlich vernachlässigbar sind.

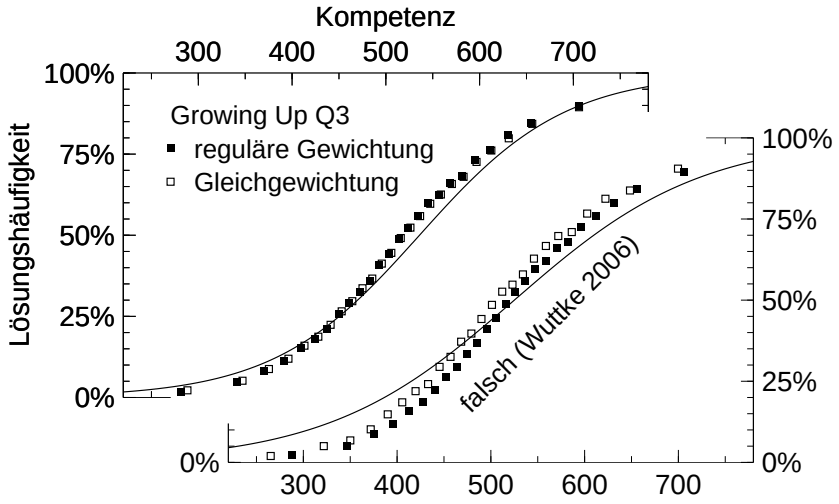

Abbildung 19: Lösungshäufigkeit als Funktion der plausiblen Kompetenzwerte, für die Aufgabe „Growing Up Q3“, für 30 gleichgewichtete OECD-Staaten. Innerhalb der einzelnen Staaten wurden die Probandengewichte einmal berücksichtigt (geschlossene Symbole), einmal nicht (offene Symbole). Die durchgezogene Kurve zeigt die Rasch-Funktion (26) zu der offiziell mitgeteilten Aufgabenschwierigkeit $\xi^{\mathrm{P}} = 574$. Im nach unten rechts verschobenen Koordinatensystem ist die fehlerhafte Auswertung aus W1 reproduziert; die Rasch-Funktion hat dort die Breite 100 statt 77,89.

D Olaf Köller

> I cannot strongly fault a 45-year-old professor for adopting this mode of defense, even though I believe it to be intellecually dishonest, because I think that for most faculty in soft psychology the full acceptance of my line of thought would involve the painful realization that one has achieved some notoriety, tenure, economic security and the like by engaging, to speak bluntly, in a bunch of nothing [Meehl 1990, S. 230].

In seinem zur politischen Schadensbegrenzung verfassten Schnellgutachten geht Köller (2006a) sehr oberflächlich über die Kritik an der Stichprobenziehung hinweg. Er bestreitet nicht die Korrelation zwischen Beteiligungsquoten und Leistungen, meint aber (Punkt 2), dieser Problematik werde Rechnung getragen, indem Länder „mit zu niedrigen Beteiligungsquoten vom Vergleich ausgeschlossen werden.“ Das ist falsch, denn die Unterschreitung vorher festgelegter Quoren blieb in mehreren Fällen folgenlos, und es ist keine adäquate Antwort auf das detailliert und quantitativ begründete Argument, dass diese Quoren nicht streng genug sind, um systematische Verzerrungen auszuschließen.

Ausführlicher äußert sich Köller zur Kritik an der Skalierung (Punkt 4):

> Wuttke kritisiert, dass die von ACER verwendete Software vermutlich nicht einmal den nationalen Projektpartnern vorgelegt wird und er auf Grund der

> Dokumentationen die Skalierung nicht nachvollziehen könne. [...] Hier erweist sich Wuttke als Laie, der weder in der Lage ist, gescheit zu recherchieren, noch die Fachdiskussion zu suchen und zu lesen. Das von ACER in PISA verwendete Softwarepaket ist Conquest, eine Skalierungssoftware, die jeder Mensch auf der Welt, auch Herr Wuttke, käuflich erwerben kann. Was Conquest genau leistet, wie man zu optimalen Schätzungen der Item- und Personenparameter kommt, ist im Conquest-Handbuch sehr gut dokumentiert. Conquest gilt aktuell als eine äußerst leistungsstarke Software, in der nur Schätzverfahren verwendet werden, die State of the Art sind. Kaufen und lesen statt zu spekulieren hätte hier Wuttke weitergeholfen.

Diese Argumentation ist aus einer ganzen Reihe von Gründen inadäquat (vgl. Meyerhöfer 2006b):

(1) Dass eine Software als leistungsstark gilt, ist ein ausgesprochen schwaches Argument gegen die Vermutung, sie könne einen Fehler enthalten – erst recht, wenn dieser Fehler, wie meine Hypothese lautete, auf den meistbeachteten Output nicht durchschlägt.

(2) Mein Verdacht lautete nicht, dass ConQuest falsch bedient wurde, sondern dass es anders rechnet, als im Technischen Bericht beschrieben. Um das zu klären, ist der käufliche Erwerb des Binärcodes ein völlig ungeeigneter Weg.

(3) Um die Auswertung des Konsortiums nachzuvollziehen, genügt es nicht, ConQuest zu erwerben. ConQuest dient im Kern dazu, ein nichtlineares Gleichungssystem zu lösen. Aber schon der vorgeschaltete Schritt, die Reduktion der PISA-Rohdaten zum Input dieses Gleichungssystems, ist nicht nachvollziehbar (3.13).

(4) Kommerzialisierung von Software ist ziemlich genau das Gegenteil von Offenlegung. Dass sich ACER für den Binärcode bezahlen lässt, bestätigt meine Vermutung, dass die nationalen Projektpartnern den Quelltext nicht vorgelegt bekommen, den implementierten Algorithmus also ebensowenig wie ich im Detail überprüfen können.

(5) Zur Qualität der Dokumentation im ConQuest-Handbuch siehe oben (3.2).

Der andere Kritikpunkt, auf den Köller ausführlich eingeht, betrifft die Modellierung des Antwortverhaltens (Punkt 5):

> Wuttke argumentiert, dass das verwendete 1-Parameter-Rasch-Modell ungeeignet ist, um Aufgaben- und Personenparameter zu schätzen. Hier hätten Mehr-Parameter-Modelle verwendet werden müssen. Auch hier erweist sich Wuttke als Laie. Hätte er sich mit der großen Literaturmenge zu IRT-Modellen auseinandergesetzt, wäre er zu anderen Schlüssen gekommen. Bos hat die IGLU-2001-Daten, die mit dem 3-Parameter-Modell skaliert worden sind, noch einmal mit Conquest (1-Parameter-Modell) skaliert und dabei festgestellt, dass sich die Aufgaben- und Personenparameter quasi nicht unterschieden, die unterschiedlichen Modelle hatten keinen differenziellen Effekt auf das Kompetenzmodell. In der IRT-Literatur ist man sich einig, dass die verschiedenen Modelle „unter dem

> Strich“ zu weitgehend identischen Schätzungen der Personenparameter führen und es eher eine Frage der Weltanschauung ist, welches man verwendet (Europa und Australien eher das 1-Parameter-Modell, d[i]e USA eher das 2 und 3-Parameter-Modell). Das in Conquest verwendete 1-Parameter-Modell mit der Bestimmung der Plausible Values hat den großen Vorteil, dass es die besten Schätzungen für den Mittelwert und die Varianz eines Landes liefert. [...]

Schon die Zitierweise („Bos“) deutet nicht gerade darauf hin, dass sich Köller auf eine wissenschaftliche Auseinandersetzung einlassen möchte. Er zitiert mich so, als hätte ich unqualifiziert *behauptet*, dass „das 1-Parameter-Raschmodell ungeeignet ist, um Item- und Personenparameter zu schätzen“, und das zum *Ausgangspunkt* einer Argumentation gemacht; er übergeht vollständig, dass ich anhand des empirischen Datenmaterials *bewiesen* habe, dass das Raschmodell bei manchen Items eklatant falsche Parameter liefert.

Unklar bleibt, warum sich Bos, indem er die Validität verschiedener Modelle empirisch überprüft hat, nicht auch als Laie erwiesen hat. Selbst wenn das Rasch-Modell bei IGLU tatsächlich funktioniert haben sollte, bewiese das für PISA überhaupt nichts.

Köllers Bemerkung, in der Literatur sei man sich einig, dass verschiedene Modelle unter dem Strich zu „weitgehend identischen“ Schätzungen der Personenparameter führen, suggeriert, ich hätte Gegenteiliges behauptet. Das ist falsch. In W1 (S. 149) habe ich vielmehr ausdrücklich darauf hingewiesen, es sei *nicht* zu erwarten, dass eine Neuauswertung die bisherigen Ranglisten auf den Kopf stellen werde. Die Aufgabenparameter aber, die Köller an dieser Stelle nicht mehr erwähnt, hängen *sehr* von der Modellierung des Antwortverhaltens ab.

Wenn in der IRT-Literatur überhaupt Einigkeit besteht, dann, dass das Rasch-Modell in einer Vielzahl von Situationen unangemessen ist. In seiner „Brief History of Item Response Theory“ bewertet es Bock (1997, S. 27) so:

> Although this solution to the item-parameter estimation problem is of interest theoretically, it does not satisfy the requirements of practical testing programs. [...] It also assumes the item slopes to be equal when more often in practical testing they are unequal. As a result, there is no possibility of estimating item discriminating powers, which are essential in test construction for choosing items that ensure good test reliability [...]

Kubinger (2000) warnt, dass

> formal wie inhaltlich zu wenig sorgfältig konzipierte Tests bei einer Prüfung dem Rasch-Modell eben nicht standhalten, sie also mit ihren Aufgaben keine faire Verrechnung der Testleistungen bieten.

Rost (1999), Koautor der deutschen PISA-Berichte, argumentiert,

> daß sich der praktische Nutzen der Rasch-Meßtheorie erst entfaltet, wenn man die Ebene des einfachen dichotomen Rasch-Modells verläßt und die zahlreichen Verallgemeinerungen dieses Modellansatzes einbezieht.

Von Weltanschauung ist nirgendwo die Rede; es scheint eher, dass die Wahl des Antwortmodells eine Frage des Erkenntnisinteresses und der Sorgfalt ist.

Der letzte Satz aus der zitierten Gutachten-Passage ist entlarvend: Köller begründet die in PISA gewählte Auswertemethodik nicht mathematisch, sondern mit dem Leistungsumfang verfügbarer Software.[76] Aus 3.6 ff. ist leicht ersichtlich, dass eine Maximum-Likelihood-Schätzung der Aufgabenparameter und eine Ziehung plausibler Kompetenzwerte sich ohne weiteres auch für mehrparametrige Antwortmodelle implementieren lässt.

E Die deutsche PISA-Expertengruppe Mathematik

Unter dem Kollaborationsnamen „Deutsche PISA-Expertengruppe Mathematik, PISA-2000“ haben Knoche und acht Koautoren (2002) einen umfangreichen Bericht über den internationalen und nationalen Mathematiktest veröffentlicht, der vielversprechend beginnt:

> Der folgende Beitrag stellt die Konzeptionen beider Tests und Analysen der Ergebnisse vor. Dabei wird in die Betrachtungen auch eine Darstellung der messtheoretischen Verfahren, die in die Konzeptionen der Tests wie in die Analysen eingehen, so weit aufgenommen, dass der Leser die vorgestellten Analysen mit Blick auf beide Komponenten – die Konzeption und das Analyseverfahren – selbst nachvollziehen kann.

Dies war einer der ersten Texte, die ich zu PISA gelesen habe. Ich sah mich damals außerstande, die vorgestellten Analyseverfahren nachzuvollziehen. Zufällig bin ich jetzt, nach gründlicher Beschäftigung mit der Item-Response-Methodik und der PISA-Auswertung, noch einmal auf jenen Bericht aufmerksam geworden – und kann nun sehr genau benennen, warum er auf Nichtspezialisten unzugänglich wirken *muss*, und dass es weder mit dem Expertentum, noch erst recht mit dem didaktischen Können der Autoren weit her ist. Dazu werde ich die Schlüsselpassage, in der die Parameterschätzung behandelt wird, in extenso zitieren und analysieren.

Teil II des Aufsatzes steht unter der Überschrift „Methodische Aspekte der Testkonzeption“. Die Skalierung der Daten wird in II.1 „Modellbetrachtungen und Skalen“ auf sieben Seiten abgehandelt. Zu Beginn (S. 165 f.) wird die Datenstruktur für den Fall dichotomer Aufgaben eingeführt, einiges an Notation

[76] Eine am Max-Planck-Institut für Bildungsforschung angefertigte und von Köller mitbegutachtete Doktorarbeit (Brunner 2005, S. 136) bestätigt, dass deutsche PISA-Experten in ihren Auswertemöglichkeiten durch die vorhandene Software begrenzt sind und im Bedarfsfall nicht einmal erwägen, eine Item-Response-Parameterschätzung selbst zu programmieren.

festgelegt und das Antwortmodell von Rasch postuliert. Es wird erläutert, dass Lösungswahrscheinlichkeiten nur von Differenzenen $\delta - \theta$ abhängen (hier ist δ die Aufgabenschwierigkeit, bei mir ξ), und dass daher bei der Schätzung der Modellparameter eine Zwangsbedingung erforderlich ist, um die Skalen zu fixieren. Sodann wird eine zweite Schreibweise für das Rasch-Modell eingeführt und die Lösungswahrscheinlichkeit für verschiedene δ als Funktion von θ aufgetragen.

Die Erklärung der Skalierung wird durch einen Exkurs über Multiple-Choice-Aufgaben unterbrochen: „an sich“ sei das Rasch-Modell unangemessen; durch geeignete Aufgabenkonstruktion lasse sich der Rateeffekt aber „soweit abschwächen, dass die Schätzung der Schwierigkeitsparameter kaum noch verzerrt wird.“ Dieser Exkurs ist länger als die nun folgende Beschreibung der Parameterschätzung (S. 167f.):

> Bei der Modellschätzung mit CONQUEST wird die theoretische Wahrscheinlichkeit L der beobachteten Datenmatrix unter der Annahme maximiert, dass die Personenparameter approximativ nach einem vorgegebenen Verteilungstyp (zum Beispiel einer Normalverteilung mit $\mu_\theta = 0$ und einer zu schätzenden Standardabweichung σ_θ) verteilt sind. Man nennt L die *Likelihood* der Datenmatrix. Dabei definiert jedes Testheft als Teiltest seine eigenen Schätzgleichungen und es wird als Nebenbedingung bei der Schätzung verlangt, dass der Schätzwert δ_a für den Schwierigkeitsparameter einer *Ankeraufgabe* in allen Teiltests der gleiche ist, in denen a vorkommt.
>
> Da auf der Probandenseite unabhängig von der Populationsgröße nur wenige Verteilungsparameter zu schätzen sind, sind die Schätzungen der Schwierigkeitsparameter konsistent.
>
> Werden für Detailanalysen auch Fähigkeitsparameter von Einzelpersonen gebraucht, so kann für einen Probanden mit dem Antwortvektor $(x_1, \ldots, x_n) \in \{0; 1\}^{\times n}$ zu n Aufgaben mit den vorab mit CONQUEST geschätzten Schwierigkeitsparametern $\delta_1, \ldots, \delta_n$ der Fähigkeitsparameter θ nach der Maximum-Likelihood-Methode durch Lösen der folgenden Gleichung geschätzt werden:
>
> (II.1.2) $$\sum_{i=1}^{n} \frac{1}{1 + \exp(\delta_i - \theta)} = \sum_{i=1}^{n} x_i \,.$$
>
> Dieses Verfahren (kurz *ML-Schätzung* genannt) hat den Nachteil, dass Probanden mit 0 oder n richtig bearbeiteten Aufgaben kein Schätzwert für θ zugewiesen werden kann. So sind nur $n-1$ interpretierbare Werte für θ möglich. Außerdem muss der Schätzfehler als relativ groß angesehen werden, da die bei der Schätzung verwendete Aufgabenzahl n jeweils nur die des bearbeiteten Testhefts ist. Für die PISA-Studie wurde daher wie schon in den TIMS-Studien das Verfahren der *Plausible Values* verwendet.

Eine „Datenmatrix“ (Zeilen 2, 5) ist zuvor nicht eingeführt worden; sie ergibt sich auch nicht einfach durch Zusammenfassen der „Antwortvektoren“ (Z. 14) – siehe oben Fußnote 25.

In Z. 4 beschreiben Knoche *et al.* plötzlich nicht mehr, wie in PISA ausgewertet wurde, sondern wie „zum Beispiel“ ausgewertet worden sein könnte – genau

wie Adams im Technischen Bericht, der über weite Strecken nicht referiert, was konkret gerechnet wurde, sondern was man mit ConQuest alles rechnen *kann* (3.2). Zwar haben Knoche *et al.* das zutreffende Beispiel gewählt: in PISA werden die Probandenkompetenzen als normalverteilt mit fixem Mittelwert und anzupassender Breite angenommen (3.5, Gl. 10). Aber in Z. 10 zeigt sich, dass sie zuvor nur zufällig richtig geraten haben: es sind nicht „wenige“ Verteilungsparameter zu schätzen, sondern exakt einer.[77]

Um nachvollziehen zu können, wie L maximiert wird, müsste man wissen, welche Parameter dabei variiert werden. Laut Z. 5 die Standardabweichung der Personenparameterverteilung (hier σ_θ, bei mir zwecks Unterscheidung von a-posteriori-Standardabweichungen δ). Tatsächlich aber auch und vor allem die Schwierigkeitsparameter.

Z. 6–9 besagt nicht mehr, als dass bei der Parameterschätzung die Verteilung der Aufgaben auf verschiedene Testhefte berücksichtigt wird. Ob dies in Form von separaten, aber durch Nebenbedingungen verknüpften Gleichungen geschieht, ist ein völlig irrelevantes technisches Detail – zumindest, solange die Gleichungen selbst nicht mitgeteilt werden.

Der nächste Satz und Absatz (Z. 10–12) zeigt exemplarisch, warum man sich als Leser dieses Textes dumm vorkommen muss. Vorausgesetzt wird die Kenntnis des recht speziellen Fachausdrucks „konsistent“.[78] Behauptet wird, die Konsistenz einer bestimmten Schätzung folge aus der geringen Anzahl bestimmter Parameter. Suggeriert wird durch das Fehlen eines Literaturhinweises, diesen Schluss müsse der typische Leser des Journals für Mathematik-Didaktik nachvollziehen können. Es gibt aber kein allgemeines Theorem, das einen Schluss von *wenige Parameter der einen Sorte* auf *konsistente Schätzung von Parametern der anderen Sorte* erlaubt. Ein solcher Schluss kann immer nur für bestimmte Modelle oder Klassen von Modellen begründet werden. In einem technisch detaillierten Buch über Rasch-Modelle (Molenaar in Fischer/Molenaar S. 46) werden solche Theoreme erwähnt, aber nicht mit allen ihren Voraussetzungen abgedruckt, geschweige denn bewiesen: dafür wird auf eine Doktorarbeit und einen Konferenzbericht verwiesen, denn die Frage der Konsistenz ist selbst für Item-Response-Spezialisten ein Seitenschauplatz von minimaler praktischer Bedeutung (*loc. cit.*, S. 47, 49). *Dass* Knoche *et al.* diesen Punkt in einer so kurzen Darstellung der Skalierung überhaupt berühren, zeugt von wenig Urteilskraft;

[77] Einer pro Testgebiet, aber die Zerlegung des Tests nach inhaltlichen Gebieten haben Knoche *et al.* bis hierhin nicht erwähnt.

[78] Es sei ein Modell gegeben, in dem beobachtbare Größen von bestimmten Parametern abhängen. Eine *Schätzung* ist der Versuch, von empirischen Beobachtungen auf die zugrunde liegenden Parameter zu schließen. *Konsistent* heißt eine Parameterschätzung, wenn sie bei Zunahme des Beobachtungsmaterials (Stichprobenumfang $\to \infty$) gegen die vorausgesetzten Modellparameter konvergiert. Dass sich Konsistenz nicht von selbst versteht, zeigen gezielt konstruierte Gegenbeispiele (Romano/Siegel 1986, S. 225).

wie sie ihn erwähnen, von Herablassung gegenüber dem als Nichtspezialist anzunehmenden Leser.

Nachdem Knoche *et al.* solcherart die Schätzung der Aufgabenschwierigkeiten und Kompetenzverteilungsparameter abgehandelt haben, wenden sie sich den individuellen Kompetenzwerten zu. Der einleitende Nebensatz (Z. 13 f.) ist irreführend: „Fähigkeitsparameter von Einzelpersonen“ werden nicht nur für irgendwelche „Detailanalysen“ gebraucht, sondern für jede Aussage über Kompetenzen von Subpopulationen; die Bestimmung individueller Kompetenzen ist integraler Bestandteil der PISA-Auswertung, auch wenn man sich letztlich nie für Individuen interessiert, sondern immer über viele Probanden mittelt.

Die Maximum-Likelihood-Gleichung in Z. 18 ist wohl als didaktischer Umweg gemeint, denn im folgenden wird erklärt, dass in PISA an ihrer statt das plausible-Werte-Verfahren angewandt wird. Dabei geht einiges durcheinander. Die angegebene Gleichung ist im PISA-Kontext falsch: sie berücksichtigt nicht die latente Verteilung der Kompetenzwerte. Genau diese latente Verteilung behebt auch das Problem, Probanden mit 0 oder n richtig bearbeiteten Aufgaben kein θ zuweisen zu können (Z. 19–21); ob man die Kompetenzen über wahrscheinlichste Werte (maximum likelihood) oder über volle Wahrscheinlichkeitsdichten (kommuniziert über plausible Werte) ausdrückt, hat damit überhaupt nichts zu tun.

Zusammengefasst: Die deutschen PISA-Mathematik-Experten werden ihrem Anspruch, eine nachvollziehbare Darstellung der messtheoretischen Verfahren zu geben, nicht gerecht. Unwissen über das verwendete Modell, Auslassungen, Ungenauigkeiten und Fehler in zentralen Punkten, Überbetonung von Seitenargumenten und die Berufung auf Möglichkeiten einer bestimmten Software deuten weder auf souveräne Beherrschung der Theorie, noch auf präzise Kenntnis der konkret in PISA angewandten Prozeduren. Als Didaktiker erweisen sich die Autoren erst recht nicht.

Siglen

D00 = Baumert *et al.* (2001), PISA 2000.
D03a = Prenzel *et al.* (2004a), PISA 2003 [Kurzfassung].
D03b = Prenzel *et al.* (2004b), PISA 2003 [Langfassung].
DAM = OECD (2005b) [Data Analysis Manual].
LTW = OECD (2004a) [Learning for Tomorrow's World].
TR = OECD (2005a) [Technical Report].
W1 = Wuttke (2006) [Dieses Buch, erste Auflage].

Literatur

ACER (2005): PISA 2003 International Database. Online-Resource `http://pisaweb.acer.edu.au/oecd_2003/oecd_pisa_data_s1.html`, Datenstand vom 9. 11. 05.

Adams, R. J. / Wilson, M. / Wu, M. (1997a): Multilevel Item Response Models: An Approach to Errors in Variables Regression. J. Educ. Behav. Stat. 22 (1) 47–76.

Adams, R. J. / Wilson, M. / Wang, W.-C. (1997b): The Multinomial Random Coefficients Multinomial Logit Model. Appl. Psych. Meas. 21 (1) 1–23.

Adams, R. / Wu, M. (Hrsg.) (2002): PISA 2000 Technical Report. Paris: OECD.

Adams, R. J. (2003): Response to „Cautions on OECD's Recent Educational Survey (PISA)". Oxford Rev. Educ. 29 (3) 377–389.

Aebli, H. ([9]1976): Grundformen des Lernens. Eine Allgemeine Didaktik auf kognitionspsychologischer Grundlage. Stuttgart: Klett.

Altman, D. G. *et al.* (2001): The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration. Ann. Int. Med. 134 (8) 663–694.

Andersen, E. B. (1977): Sufficient Statistics and Latent Trait Models. Psychometrika 42 (1) 69–81.

Artelt, C. / Baumert, J. (2004): Zur Vergleichbarkeit von Schülerleistungen bei Leseaufgaben unterschiedlichen sprachlichen Ursprungs. Z. Pädagog. Psychol. 18 (3/4) 171–185.

Baker, F. B. (1992): Item Response Theory. Parameter Estimation Techniques. New York: Marcel Dekker.

Baumert, J. / Klieme E. / Lehrke, M. / Savelsbergh, E. (2000): Konzeption und Aussagekraft der TIMSS-Leistungstests. Zur Diskussion um TIMSS-Aufgaben aus der Mittelstufenphysik. Die Deutsche Schule 92 (1) 102– 115.

Baumert, J. *et al.* (Deutsches PISA–Konsortium) (2001): PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Opladen: Leske + Budrich. Zitiert als D00.

Baumert, J. / Artelt, C. (2005): Viel Lärm um nichts? — Ein Kommentar zu Eberhard Schröders gleichnamigem Forschungsreferat auf der DGPs-Tagung 2004. `http://www.uni-saarland.de/fak5/ezw/fg_paedpsych/newsletterarchiv/newsletter_1_2005/Debatte_PISA.pdf` [8. 8. 07].

Baumert, J. / Brunner, M. / Lüdtke, O. / Trautwein, U. (2007): Was messen internationale Schulleistungsstudien? — Resultate kumulativer Wissenserwerbsprozesse. Psychol. Rundsch. 58 (2) 118–128.

Bender, P. (2005): PISA, Kompetenzstufen und Mathematik-Didaktik. J. Math.-did. 26 (3/4) 274–281.

Bender, P. ([1]s2006): Was sagen uns Pisa & Co., wenn wir uns auf sie einlassen? Dieses Buch.

Bender, P. (2007): Weitere Anmerkungen zu PISA, zu PISA-Reaktionen und Reaktionen auf PISA-Reaktionen. Mitteilungen der GDM [Ges. f. Did. d. Math.] 83 (im Druck).

Bethge, T. (1999): Zum Umgang mit den Ergebnissen von TIMSS. Die Deutsche Schule 91 (2) 178–181.

Blanke, I. / Böhm, B. / Lanners, M. (2004): Beispielaufgaben und Schülerantworten. Le Gouvernement du Grand-Duchè de Luxembourg. Ministère de l'Èducation nationale et de la Formation professionelle.

Bloxom, B. / Pashley, P. J. / Nicewander, W. A. / Yan, D. (1995): Linking to a Large-Scale Assessment: An Empirical Evaluation. J. Educ. Behav. Stat. 20 (1) 1–26.

Blum, A. / Guèrin-Pace, F. (2000): De Lettres et des Chiffres. Des tests d'intelligence à l'èvaluation du „savoir lire", un siècle de polèmiques. Paris: Fayard.

Bock, R. D. / Aitkin, M. (1981): Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm. Psychometrika 46 (4) 443–459 nebst Erratum 47 (3) 369.

Bock, R. D. (1997): A Brief History of Item Response Theory. Educational Measurement: Issues and Practice 16 (4) 21–33.

Boe, E. E. / May, H. / Boruch, R. F. (2002): Student Task Persistence in the Third International Mathematics and Science Study: A Major Source of Achievement Differences at the National, Classroom, and Student Levels. Report CRESP-RR-2002-TIMSS1. Philadelphia: Pennsylvania University. `http://www.gse.upenn.edu/cresp/pdfs/20070130151136207.pdf` [8. 8. 07].

Bonnet, G. (2002): Reflections in a Critical Eye: on the pitfalls of international assessment. Assessment in Educ. 9 (3) 387–399.

Bottani, N. / Vrignaud, P. (2005): La France et les évaluations internationales. Rapport établi à la demande du Haut Conseil de l'évaluation de l'école. Online-Resource `http://lesrapports.la documentation francaise.fr/BRP/054000359/0000.pdf` [8. 8. 07].

Brügelmann, H. (2006): Tests statt Noten? Warum PISA, VERA & Co kein Modell für die Leistungsbeurteilung von SchülerInnen sein können. Vortrag auf dem Symposium „Aus PISA gelernt?", Berlin, 24. 11. 2006.

Brunell, V. (2004): Utmärkta PISA-resultat också i Svenskfinland. Pressemitteilung des Pedagogiska Forskningsinstitutet, Jyväskylä Universitet. `http://ktl.jyu.fi/pisa/Langt_pressmeddelande.pdf` [9. 8. 07].

Brunner, M. (2005): Mathematische Schülerleistung — Struktur, Schulformunterschiede und Validität. Dissertation, Humboldt-Universität Berlin.

Buckheit, J. B. / Donoho, D. L. (1995): WaveLab and Reproducible Research. In Antoniadis, A. / Oppenheim, G. (eds.): Wavelets and Statistics. New York: Springer.

Burba, D. (2006): Leistungen bei Jungen und Mädchen bei PISA 2003 — bedeutsame Unterschiede? Dissertation, Christian-Albrechts-Universität Kiel.

Carroll, J. B. (1987): The National Assessments in Reading: Are We Misreading the Findings? Phi Delta Kappan 68, 424–430.

CEPED [Centre français sur la population et le développement] (2006): Le déficit des femmes en Asie: Tendances et perspectives. Chronique no. 51.

von Collani, E. (2001): OECD PISA - An Example of Stochastic Illiteracy? Economic Quality Control 16 (2) 227–253.

Cook, L. L. / Eignor, D. R. / Taft, H. L. (1988): A Comparative Study of the Effects of Recency of Instruction on the Stability of IRT and Conventional Item Parameter Estimates. J. Educ. Meas. 25 (1) 31–45.

DESCO [Direction générale de l'Enseignement scolaire, Ministère de l'Éducation nationale, Frankreich] (2003): Évaluation des connaissances et des compétences des élèves de 15 ans: questions et hypothèses formulées à partir de l'étude de l'OCDE. Rencontres de la DESCO, 31 mai 2002. `http://eduscol.education.fr/D0122/evaluation_accueil.htm` [24. 8. 07].

Ebenrett, H. J. / Hansen, D. / Puzicha, K. J. (2003): Verlust von Humankapital in Regionen mit hoher Arbeitslosigkeit. Aus Politik u. Zeitgesch. B 06-07, 25-31.

Fischer, G. H. / Molenaar, I. W. (1995): Rasch Models. Foundations, Recent Developments, and Applications. New York: Springer.

Frederiksen, N. / Mislevy, R. J. / Bejar, I. I. (Hrsg.) (1993): Test Theory for a New Generation of Tests. Hillsdale: Lawrence Erlbaum.

Freudenthal, H. (1975): Pupils achievements internationally compared — the IEA. Educ. Stud. Math. 6, 127–186.

Ganzeboom, H. B. G. / De Graaf P. M. / Treiman, D. J. (1992): A Standard International Socio-Economic Index of Occupational Status. Soc. Sci. Res. 21 (1) 1–56.

Gellert, U. ([1]2006): Mathematik „in der Welt" und mathematische „Grundbildung". Zur Konsistenz des mathematikdidaktischen Rahmens von PISA. Dieses Buch.

Gentleman, R. C. *et al.* (2004): Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5:R80.

Gill, J. (1999): The Insignificance of Null Hypothesis Significance Testing. Polit. Res. Quart. 52 (3) 647—674.

Glas, C. A. W. (2005): Book Review [de Boeck, P. / Wilson, M. eds. (2004): Explanatory Item response Models: A Generalized Linear and Nonlinear Approach]. J. Educ. Meas. 42 (3) 303–307.

Goldstein, H. (2004): International comparisons of student attainment: somme issues arising from the PISA study. Assessment Educ. 11 (3) 319–330.

Grabe, M. (2000): Gedanken zur Revision der Gauß'schen Fehlerrechnung. Tech. Mess. 67 (6), 283–288.

Gräber, W. / Stork, H. (1984): Die Entwicklungspsychologie Jean Piagets als Mahnerin und Helferin des Lehrers im naturwissenschaftlichen Unterricht. Teil 1. MNU [Der math. naturw. Unt.] 37 (4) 193–201.

Guttman, L. (1954): Some Necessary Conditions for Common-Factor Analysis. Psychometrika 19 (2) 149–161.

Hagemeister, V. (1999): Was wurde bei TIMSS erhoben? Über die empirische Basis einer aufregenden Studie. Die Deutsche Schule 91 (2) 160–177.

Hagemeister, V. ([1]2006). Kritische Anmerkungen zum Umgang mit den Ergebnissen von PISA. Dieses Buch.

Haider, G. (Hrsg.) (2001): PISA 2000. Technischer Report. Innsbruck: StudienVerlag. `http://www.pisa-austria.at/pisa2000/index2.htm`.

Haladyna, T. M. /Nolen, S. B./Haas, N. S. (1991): Raising Standardized Achievement Test Scores and the Origins of Test Score Pollution. Educ. Researcher 20 (5) 2–7.

Hambleton, R. K. / Swaminathan, H. / Rogers, H. J. (1991): Fundamentals of Item Response Theory. Newbury Park: Sage.

Hardwig, J. (2006): Epistemic Dependence. In: Selinger, E./Crease, R. P. (Hrsg.): The Philosophy of Expertise. New York: Columbia University Press.

Hothorn, T. (2006): Praktische Aspekte der Reproduzierbarkeit statistischer Analysen in klinischen Studien, Antrittsvorlesung, Friedrich-Alexander-Universität Erlangen-Nürnberg. `http://www.imbe.med.uni-erlangen.de/~hothorn/talks/AV.pdf` [8. 8. 07].

Institut für Demoskopie Allensbach (2005). Ärzte vorn. Allensbacher Berufsprestige-Skala 2005. allensbacher berichte 2005/12.

ISO [International Organization for Standardization] (1995): Guide to the Expression of Uncertainty in Measurement. Genf: ISO.

Jablonka, E. ([1]2006): Mathematical Literacy: Die Verflüchtigung eines ambitionierten Test-Konstrukts in bedeutungslose PISA-Punkte. Dieses Buch.

Keitel, C. ([1]2006): Der (un)heimliche Einfluss der Testideologie. Dieses Buch.

Kessel, W. (2001): Der ISO/BIPM-Leitfaden zur Ermittlung der Messunsicherheit. Tech. Mess. 68 (1) 5–13.

Kießwetter, K. (2002): Unzulänglich vermessen und vermessen unzulänglich: PISA u. Co. Mitt. Dtsch. Math.-Ver. (4) 49–58.

Kim, D.-S. (2004): Le déficit de filles en Corée du Sud: évolution, niveaux et variations régionales. Population [Paris] 59, 982–997.

Kirsch, I. / de Jong, J. / Lafontaine, D. / McQueen, J. / Mendelovits, J. / Monseur, C. (2002): Reading for Change. Performance and Engagement Across Countries. Results from PISA 2000. Paris: OECD.

Klemm, K. (2006): Fünf Jahre nach dem PISA-Schock. Interview mit WDR.de. `http://www.wdr.de/themen/kultur/bildung_und_erziehung/brennpunkt_schule/pisa_co/pisa_5jahre/index.jhtml` [8. 8. 07].

Knoche, N. / Lind, D. / Blum, W. / Cohors-Fresenborg, E. / Flade, L. / Löding, W. / Möller, G. / Neubrand, M. / Wynands, A. (Deutsche PISA-Expertengruppe Mathematik, PISA-2000) (2002): Die PISA-2000-Studie, einige Ergebnisse und Analysen. J. Math.-did. 23 (3/4) 159–202.

Kohn, A. (2000): The Case Against Standardized Testing. Raising the Scores, Ruining the Schools. Portsmouth NH: Heinemann.

Köller, O. (2006a): Stellungnahme zum Text von Joachim Wuttke: „Fehler, [...]“ `http://www.iqb.hu-berlin.de/aktuell?pg=a_3` [8. 12. 06; dort inzwischen gelöscht]; `http://www.math.uni-potsdam.de/prof/o_didaktik/pisa_debatte/koeller.pdf` [8. 8. 07].

Köller, O. (2006b): Kritik an PISA ist unberechtigt. Interwiev. Bildungsklick: `http://bildungsklick.de/a/50155/kritik-an-pisa-unberechtigt` [8. 8. 07].

Kraus, J. (2005). Der PISA-Schwindel. Wien: Signum.

Kubinger, K. D. (2000): Replik auf Jürgen Rost „Was ist aus dem Rasch-Modell geworden?“: Und für die Psychologische Diagnostik hat es doch revolutionäre Bedeutung. Psychol. Rundsch. 51 (1) 33–34.

Lind, D. (2004): Welches Raten ist unerwünscht? Eine Erwiderung. J. Math.-did. 25 (1) 70–74.

Lind, D. / Knoche, N. / Blum, W. / Neubrand, M. (2005): Kompetenzstufen in PISA. – eine Erwiderung auf den Beitrag von W. Meyerhöfer [...] J. Math.-did. 25 (1) 80–87.

Lind, G. (2005): Sind die PISA-Daten in Bayern verzerrt? `http://forum-kritische-paedagogik.de/start/request.php?124` [8. 8. 07].

Luhmann, N. ([4]1974): Selbststeuerung der Wissenschaft. In: Soziologische Aufklärung. Aufsätze zur Theorie sozialer Systeme. Band 1. Opladen: Westdeutscher Verlag [zuerst in: Jahrb. Sozialwiss. 19, 147–170 (1968)].

Mahamed, A. / Gregory, P. A. M. / Austin, Z. (2006): “Testwiseness” Among International Pharmacy Graduates and Canadian Senior Pharmacy Students. Am. J. Pharm. Educ. 70 (6) 131.

Martin, M. O. / Kelly, D. L. (Hrsg.) (1998): Third International Mathematics and Science Study Technical Report, Volume III: Implementation and Analysis — Final Year of

Secondary School. Chestnut Hill: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Mathis, W. J. (2004): NCLB and High-Stakes Accountability: A Cure? Or a Symptom of the Desease? Educ. Horizons 82 (2) 143–152.

McCluskey, H. / Zahner Rossier, C. (2004): Das Projekt PISA und die Durchführung in der Schweiz. Bundesamt für Statistik / Eidgenössische Konferenz der kantonalen Erziehungsdirektoren. `http://www.portal-stat.admin.ch/pisa/download/pisa_description_d.pdf` [8. 1. 07, am 8. 8. 07 dort nicht mehr auffindbar].

McLachlan, G. J. / Krishnan, T. (1997): The EM Algorithm and Extensions. New York: Wiley.

Meehl, P. E. (1978): Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology. J. Consult. Clin. Psychol. 46 (4) 806–834.

Meehl, P. E. (1990): Why summaries of research on psychological theories are often uninterpretable. Psycholog. Rep. 66, 195–244.

Meyerhöfer, W. (2004a): Zum Problem des Ratens bei PISA. J. Math.-did. 25 (1) 62–69.

Meyerhöfer, W. (2004b): Zum Kompetenzstufenmodell von PISA. J. Math.-did. 25 (3/4) 294–305.

Meyerhöfer, W. (2005): Tests im Test: Das Beispiel PISA. Leverkusen: Barbara Budrich.

Meyerhöfer, W. ([1]2006a): PISA & Co. als kulturindustrielle Phänomene. Dieses Buch.

Meyerhöfer, W. (2006b): Zur Stellungnahme von Prof. Olaf Köller vom 14.11.2006 zum Text von Joachim Wuttke: Fehler, Verzerrungen, Unsicherheiten in der PISA-Auswertung. `http://www.math.uni-potsdam.de/prof/o_didaktik/pisa_debatte` [8. 8. 07].

Micceri, T. (1989): The Unicorn, the Normal Curve, and other Improbable Creatures. Psychol. Bull. 105 (1) 156–166.

Millman, J. / Bishop, C. H. / Ebel, R. (1965): An Analysis of Test-Wiseness. Educ. Psychol. Meas. 25 (3) 707–726.

Moher, D. / Altman, D. G. / Schulz, K. F. / Elbourne, D. R. (2004): Opportunities and challenges for improving the quality of reporting clinical research: CONSORT and beyond. Can. Med. Assoc. J. 171 (4) 349–350.

Monseur, C. / Wu, M. (2002): Imputation for Student Nonresponse in Educational Achievement Surveys. The International Conference on Improving Surveys, Kopenhagen, 25.–28. 8. 2002. `http://www.icis.dk/ICIS_papers/E2_5_2.pdf` [4. 5. 07].

NCES [National Center for Educational Statistics] (2006): Homeschooling in the United States: 2003. Statistical Analysis Report. Online-Resource `http://nces.ed.gov/pubs2006/homeschool/` [8. 8. 07].

NCHS [U. S. Department of Health and Human Services, Centers for Desease Control and Prevention, National Center for Health Statistics] (2005): More Boys Born Than Girls. New Report Documents Total Gender Ratios At Birth From 1940 to 2002. Online-Resource `http://www.cdc.gov/nchs/pressroom/05facts/moreboys.htm` [8. 8. 07].

Neuwirth, E. / Ponocny, I. / Grossmann, W. (Hrsg.) (2006): PISA 2000 und PISA 2003: Vertiefende Analysen und Beiträge zur Methodik. Graz: Leykam.

Nichols, S. L. / Berliner, D. (2007): Collateral Damage. How High-Stakes Testing Corrupts America's Schools. Cambridge Mass.: Harvard Education Press.

NYSED [New York State Education Department, Elementary, Middle, Secondary and Continuing Education, Office of State Assessments] (2005): English Language Arts Grade 3 Sample Test. `http://www.emsc.nysed.gov/3-8/ela-sample/gr3-bk1.pdf` [5. 5. 07].

NYSED (2006): Grade 3–8 Mathematics Tests. `http://www.nysedregents.org/testing/mathei/06exams/gr3bk1.pdf` [8. 8. 07].

NYSED (o. J.): Test Your Testwiseness. www.emsc.nysed.gov/osa/assesspubs/pubsarch/ActivityTestYourTestwiseness.pdf [8. 8. 07].

OECD [Organisation for Economic Co-operation and Development] (1998): Fourth Meeting of the Board of Participating Countries (Paris, 6–7 July).

OECD (1999): Measuring Student Knowledge and Skills. A New Framework for Assessment. Paris: OECD.

OECD (2001): Knowledge and Skills for Life. First Results from the OECD Programme for International Student Assessment (PISA) 2000. Paris: OECD.

OECD (2003a): The PISA 2003 Assessment Framework. Mathematics, Reading, Science and Problem Solving Knowledge and Skills. Paris: OECD.

OECD (2003b): Test Administrator's Manual — PISA 2003. Paris: OECD.

OECD (2004a): Learning for Tomorrow's World. First Results from PISA 2003. Paris: OECD. Zitiert als LTW.

OECD (2004b): Problem Solving for Tomorrow's World. First Measures of Cross-Curricular Competencies from PISA 2003. Paris: OECD.

OECD (2005a): PISA 2003 Technical Report. Paris: OECD. Zitiert als TR.

OECD (2005b): PISA 2003 Data Analysis Manual. SPSS Users. Paris: OECD. Zitiert als DAM.

OECD (2005c): Longer Term Strategy of the Development of PISA. 20th meeting of the PISA Governing Board. 3–5 October, Reykjavik, Iceland. Paris: OECD.

Olsen, R. V. / Turmo, A. / Lie, S. (2001): Learning about students' knowledge and thinking in science through large-scale quantitative studies. Eur. J. Psychol. Educ. 16 (3) 403–420.

Olsen, R. V. (2005a): Achievement tests from an item perspective. An exploration of single item data from the PISA and TIMSS studies, and how such data can inform us about students' knowledge and thinking in science. Dissertation, Universität Oslo.

Olsen, R. V. (2005b): An exploration of cluster structure in scientific literacy in PISA: Evidence for a Nordic dimension? NorDiNa 1 (1) 81–94. http://www.naturfagsenteret.no/tidsskrift/Rolf%20V%20Olsen_105.pdf [8. 8. 07].

Oster, E. (2005): Hepatitis B and the Case of the Missing Women. J. Polit. Econ. 113 (6) 1163–1216.

Paris, S. G. / Lawton, T. A. / Turner, J. C. / Roth, J. L. (1991): A Developmental Perspective on Standardized Achievement Testing. Educ. Researcher 20 (5) 12–20.

Prais, S. J. (2003): Cautions on OECD's Recent Educational Survey (PISA). Oxford Rev. Educ. 29 (2) 139–163.

Prenzel, M. / Baumert, J. / Lehmann, R. / Leutner, D. / Neubrand, M. / Pekrun, R. / Rolff, H.-G. / Rost, J. / Schiefele, U. [PISA–Konsortium Deutschland] (Hrsg.) (2004a): PISA 2003. Ergebnisse des zweiten internationalen Vergleichs. Zusammenfassung. Kiel: Leibniz-Institut für die Pädagogik der Naturwissenschaften. http://pisa.ipn.uni-kiel.de/Ergebnisse_PISA_2003.pdf [8. 8. 07]. Zitiert als D03a.

Prenzel, M. / Baumert, J. / Blum, W. / Lehmann, R. / Leutner, D. / Neubrand, M. / Pekrun, R. / Rolff, H.-G. / Rost, J. / Schiefele, U. [PISA–Konsortium Deutschland] (Hrsg.) (2004b): PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland — Ergebnisse des zweiten internationalen Vergleichs. Münster: Waxmann. Zitiert als D03b.

Prenzel, M. (2004c): Zu: Analyse der veröffentlichten Chemie-Aufgaben von PISA. MNU [Der math. naturw. Unt.] 57 (6) 377–378.

Prenzel, M. (2006): Wie solide ist PISA? oder Ist die Kritik von Joachim Wuttke begründet? http://pisa.ipn.uni-kiel.de/Wie_solide_ist_PISA.pdf [8. 8. 07].

Prenzel, M. / Walter, O. (2006): Ein Programmierfehler in PISA? Joachim Wuttke hat falsch gerechnet! Anlage zu Prenzel (2006), URL wie dort.

Prenzel, M. / Walter, O. / Frey, A. (2007): PISA misst Kompetenzen. Eine Replik auf Rindermann (2006). Was messen internationale Schulleistungsstudien? Psychol. Rundsch. 58 (2) 128–136.

Putz, M. (2005): Was misst PISA? `http://www.schule.suedtirol.it/rg-bx/projekte/Zeitung/ZeitungM%C3%A4rz05.pdf` [8. 8. 07].

Putz, M. (2006): PISA: Zu schön um wahr zu sein? Liegt das Traumergebnis an Rechenfehlern? Unveröffentlicht.

Reagan-Cirincione, P. / Rohrbaugh, J. (1992): Decision Conferencing: A Unique Approach to Behavioral Aggregation of Expert Judgement. In: Wright, G. / Bolger, F. (Hrsg.): Expertise and Decision Support. New York: Plenum.

Rindermann, H. (2006): Was messen internationale Schulleistungsstudien? Schulleistungen, Schülerfähigkeiten, kognitive Fähigkeiten, Wissen oder allgemeine Intelligenz? Psychol. Rundsch. 57 (2) 69-86.

Rindermann, H. (2007a): Intelligenz, kognitive Fähigkeiten, Humankapital und Rationalität auf verschiedenen Ebenen. Psychol. Rundsch. 58 (2) 137–145.

Rindermann, H. (2007b): The g-factor of international cognitive ability comparisons. Eur. J. Personality 21, 667–706.

Rocher, T. (2003): La méthodologie des évaluations internationales de compétences. Psychologie et Psychométrie 24 (2–3) [Numéro spécial : Mesure et Éducation], 117–146.

Romainville, M. (2002): Du bon usage de PISA. La Revue Nouvelle 115 (3–4) 86–99.

Romano, J. P. / Siegel, A. F. (1986): Counterexamples in probability and statistics. Monterey: Wadsworth & Brooks/Cole.

Rost, J. (1999): Was ist aus dem Rasch-Modell geworden? Psychol. Rundsch. 50 (3) 140–156.

Rost, J. (2000): Haben ordinale Rasch-Modelle variierende Trennschärfen? Eine Antwort auf die Wiener Repliken. Psychol. Rundsch. 51 (1) 36–37.

Rost, J. (^{2}2004): Lehrbuch Testtheorie — Testkonstruktion. Bern: Hans Huber.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons.

Ruddock, G. (2006): Validation study of the PISA 2000, PISA 2003 and TIMSS-2003 International studies of pupil attainment. Nottingham: Department for Education and Skills. `http://www.dfes.gov.uk/research/data/uploadfiles/RR772.pdf` [22. 7. 07].

Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. London: Chapman & Hall.

Schmidt, H. (2003): Warum GUM? Kritische Anmerkungen zur Normdefinition der „Messunsicherheit“ und zu verzerrten „Elementarfehlermodellen“. ZfV — Zs. f. Geodäsie &c. 128 (5) 303–312.

Schulz, W. (2006): Response to paper of Joachim Wuttke. Unveröffentlicht.

Shamos, M. H. (1995): The myth of scientific literacy. New Brunswick NJ: Rutgers University Press.

Shriberg, D. / Shriberg A. B. (2006): High-Stakes Testing and Dropout Rates. Dissent Magazine. `http://www.dissentmagazine.org/article/?article=702` [8. 8. 07].

Sireci, S. G. (1997): Problems and Issues in Linking Assessments Across Languages. Educational Measurement: Issues and Practice 16 (1) 12–19.

Song, S.-Y. (1998): The Problem of Sex Ratio in Asia. S. 188–190 in Fujiki, N., Macer, D. R. J. (Hrsg.): Bioethics in Asia. Eubios Ethics Institute.

U.S. Census Bureau (2006): International Data Base. Population Pyramids. Verwendete Online-Resource `http://www.census.gov/ipc/www/idbpyr.html` [2. 1. 07]; aktualisierte Daten jetzt unter `http://www.census.gov/ipc/www/idb/pyramids.html` [8. 8. 07].

Willemen, R. J. (2002): School timetable construction: algorithms and complexity. Dissertation, Technische Universiteit Eindhoven.

Wise, S. L. / DeMars, C. E. (2005): Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions. Educ. Assessment 10 (1) 1–18.

Walberg, H. J. / Strykowski, B. F. / Rovai, E. / Hung, S. S. (1984): Exceptional Performance. Rev. Educ. Res. 54 (1) 87–112.

Woodruff, D. / Hanson, B. A. (1997): Estimation for Item response Models using the EM Algorithm for Finite Mixtures. Revised Edition. Presented at the Annual Meeting of the Psychometric Society, Gatlinburg, Tennessee. `http://www.b-a-h.com/papers/paper9701.pdf` [8. 8. 07].

Woods, C. M. / Thissen, D. (2006): Item Response Theory with Estimation of the Latent Population Distribution Using Spline-Based Densities. Psychometrika 71 (2) 281–301.

Wu, M. L. / Adams, R. J. / Wilson, M. R. (1998): ACER ConQuest. Generalised Item Response Modelling Software Manual. Melbourne: The Australian Council for Educational Research Ltd.

Wu, M. / Douglas, A. R. / Monseur, C. (2002): Issues in the Design of the Student Assessment Instrument for PISA 2000. The International Conference on Improving Surveys, Kopenhagen, 25.–28. 8. 2002. `http://www.icis.dk/ICIS_papers/E2_3_3.pdf` [4. 5. 07].

Wuttke, J. ([1]2006): Fehler, Verzerrungen, Unsicherheiten in der PISA-Auswertung. Dieses Buch, erste Auflage. Zitiert als W1.

Yamamoto, K./Mazzeo, J. (1992): Item Response Theory Scale Linking in NAEP. J. Educ. Stat. 17 (2) 155–173.

Yousfi, S. (2005): Mythen und Paradoxien der klassischen Testtheorie (I). Testlänge und Gütekriterien. Diagnostica 51 (1) 1–11.

Zabulionis, A. (2001): Similarity of Mathematics and Science Achievement of Various Nations. Educ. Policy Analysis Arch. 9 (33). `http://epaa.asu.edu/epaa/v9n33/` [8. 8. 07].

Zwick, R. (1992): Statistical and Psychometric Issues in the Measurement of Educational Achievement Trends: Examples From the National Assessment of Educational Progress. J. Educ. Stat. 17 (2) 203–218.